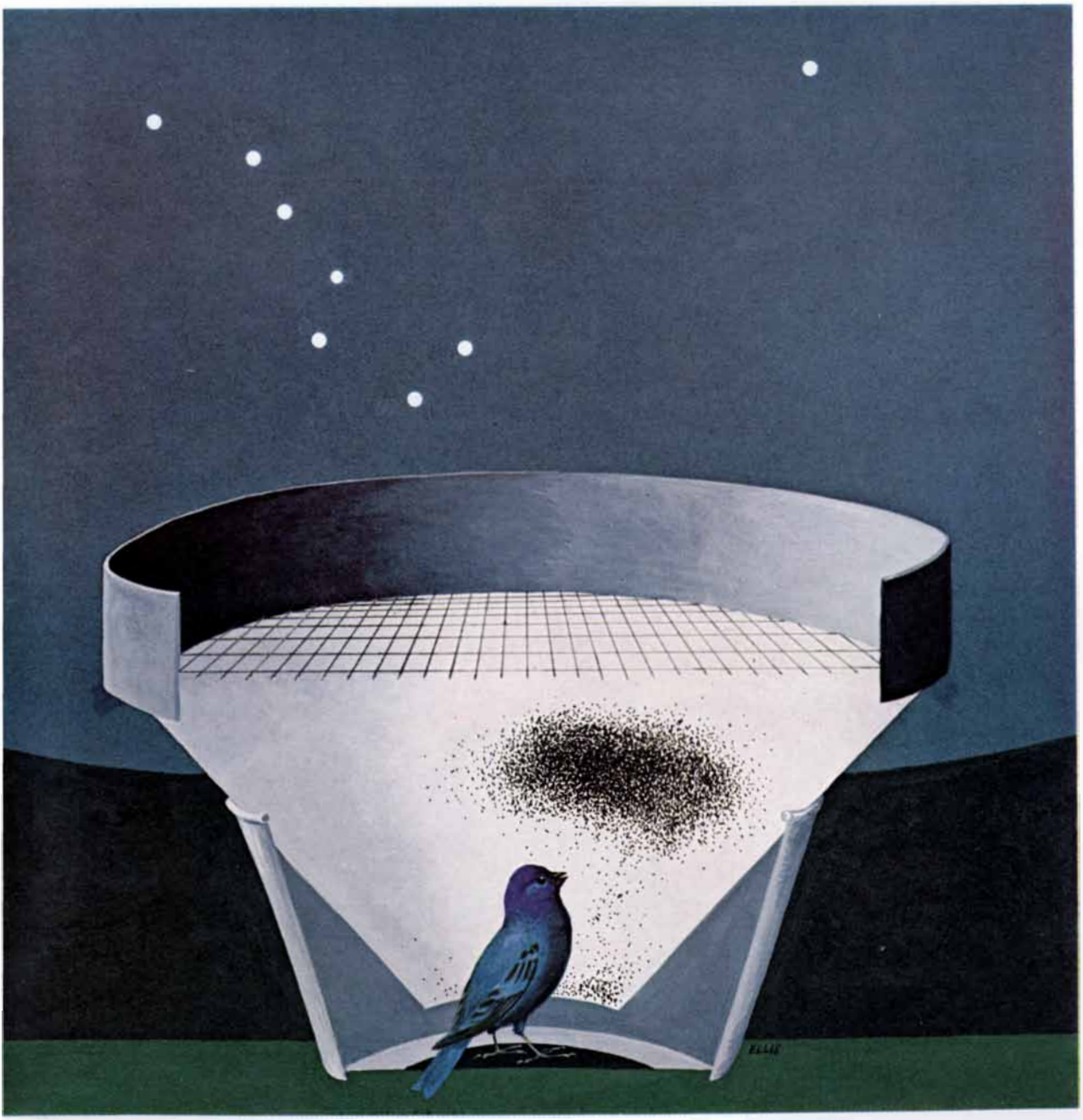
SCIENTIFIC AMERICAN



STELLAR ORIENTATION IN BIRDS

$1.25

August 1975

The way we make chlorine saves enough electricity to light the homes of Philadelphia for a year.

Chlorine is chlorine, and for the most part, one supplier's chlorine is just as good as the next guy's.

So since everybody's chlorine is basically created equal, Diamond Shamrock found a better way to make it.

Chlorine is manufactured by supplying salt brine and large amounts of electricity to electrolytic cells. At Diamond Shamrock, we design cells which utilize a Dimensionally Stable Anode (DSA*) made of exotic metals rather than a graphite anode. This Diamond Shamrock innovation allows the manufacturer of chlorine a 15% to 20% reduction in power consumption.

In fact, its use by the North American chlorine industry saves 2½ billion kilowatt-hours per year—more than enough electricity to light the homes of Philadelphia for a year.

At present, only 50% of the chlorine manufacturers use the diaphragm cell with an energy-saving Dimensionally Stable Anode. But if everyone did, think of the electricity you would save. With the energy situation being what it is, the electricity saved will have an impact on our entire national economy.

The way we make chlorine saves more than electricity.

By using the Diamond Shamrock diaphragm cell to make chlorine, not only is electricity saved, but the environment is helped.

The need to open up a Dimensionally Stable Anode Cell for cleaning is less frequent because it operates cleaner than other cells. Therefore, less contaminants are emitted to the atmosphere. Also, chlorinated organic waste is reduced 90%. Other disposable materials used up in the cells are less than half of what was expected in the past.

The way we make chlorine saves 2½ billion kilowatt-hours of electricity.

Since we make chlorine the way we do, it costs less.

Because many U.S. industries depend on Diamond Shamrock for their supply of chlorine and its co-product, caustic soda, Diamond Shamrock recently started up a 1200 ton per day chlorine/caustic soda plant near Houston, Texas—increasing our overall capacity approximately 50%.

This means availability of more chlorine for the manufacture of plastics, synthetic detergents, drugs, dyes, refrigerants, insecticides, weed killers, solvents, as a bleaching agent for paper and textiles, and for killing bacteria in water purification and sewage treatment.

The additional caustic soda will be used for the manufacture of drugs and dyes, food processing chemicals, petroleum refining, soaps and detergents, and the manufacture of aluminum.

Diamond Shamrock produces both liquid and dry caustic, in various forms and grades.

For information about how we make chlorine and caustic soda, or about Diamond Shamrock, the international corporation that's turning industries around with chemistry and energy, write: Diamond Shamrock, 1100 Superior Avenue, Cleveland, Ohio 44114.

After all, it doesn't take much energy to write.

Philadelphia

We can keep it lit for a year.

The way we make chlorine also helps the environment.

*DSA is a registered trademark of Electronor Corporation.

Diamond Shamrock

As your introduction to membership in the BOOK-OF-THE-MONTH CLUB®

The most comprehensive encyclopedia of philosophy ever published...

FOR ONLY $17.50

THE SUGGESTED TRIAL
You simply agree to buy four Club choices within a year at substantial savings on most books you choose.

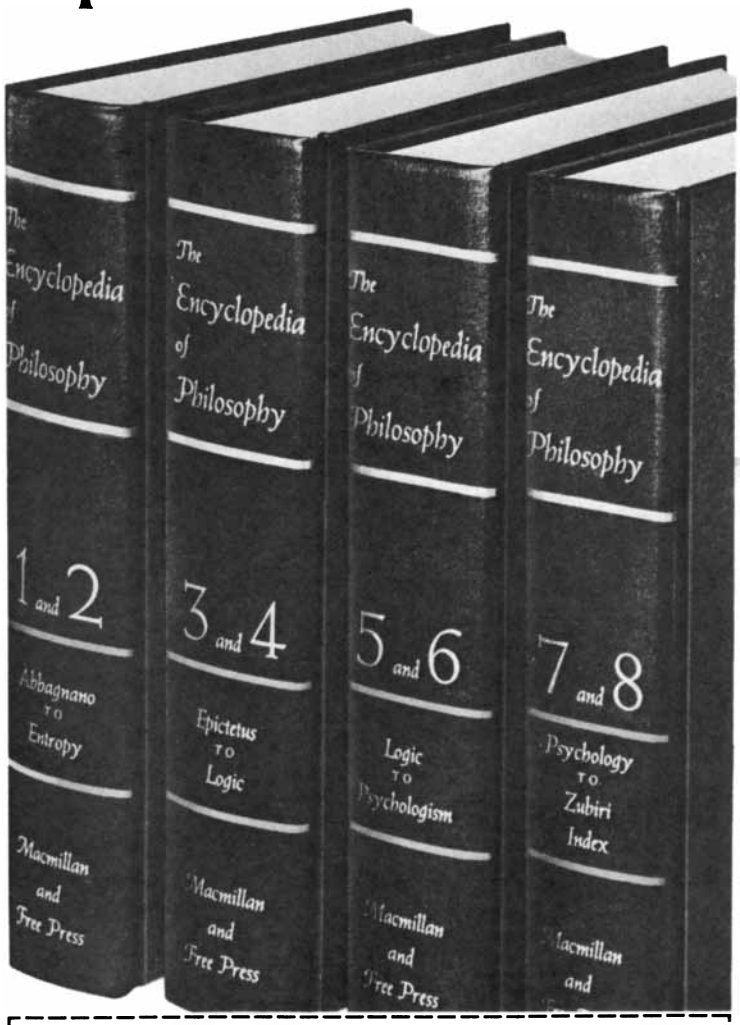
PUBLISHER'S LIST PRICE: $99.50

The complete contents of the original eight-volume $219.50 edition now in a handsome four-volume reader's edition

ONE OF THE GREATEST repositories of philosophical information available anywhere, *The Encyclopedia of Philosophy* is the only major philosophical reference work published in English since 1901.

In this unique 5,000,000-word work can be found clear, authoritative answers to almost any question concerning philosophy – ancient, medieval and modern, Eastern as well as Western. And because the work represents the collaboration of more than 500 philosophers and philosophical scholars from twenty-four countries, every shade of opinion is represented. No thinker or movement is dismissed as too radical, no issue is scanted as too controversial. There are articles, for example, that discuss such contemporary matters as the ethics of suicide and euthanasia and the benefits and dangers of computing machines.

Clearly, *The Encyclopedia of Philosophy* belongs in every personal library, particularly in homes where there are high school or college students. Up to now its high price has kept it out of the reach of most families. But, as a new member of the Book-of-the-Month Club, you may own it for less than a tenth of the cost of the original eight-volume edition. Because only a limited number of sets have been set aside for this offer, you are advised to act quickly.



SPECIAL FEATURES AND EDITORIAL FACTS

- 4200 pages
- 5,000,000 words
- Nearly 1500 articles, including 900 on individual philosophers
- 158-page index
- Paul Edwards, editor in chief
- Editorial Board of 153 international scholars
- Contributors: 500 specialists from twenty-four countries

A SMALL SAMPLING OF THE IDEAS AND PHILOSOPHERS IN THESE VOLUMES

Pain	Atheism	Semantics	The Absolute
Nothing	Free Will	Humanism	Good and Evil
Truth	Nonsense	Linguistics	Existentialism

LOCKE	WITTGENSTEIN	CAMUS	NEWTON
CONFUCIUS	SANTAYANA	KANT	DESCARTES
ERASMUS	SPINOZA	NIETZSCHE	HEGEL
SARTRE	THOMAS AQUINAS	ARISTOTLE	EINSTEIN
RUSSELL	PIAGET	PLATO	KIERKEGAARD

BOOK-OF-THE-MONTH CLUB, INC., Camp Hill, Pennsylvania, 17012 **5-A170-8**

Please enroll me as a member of the Book-of-the-Month Club and send me the four-volume reader's edition of *The Encyclopedia of Philosophy*, billing me $17.50. I agree to buy at least four Selections or Alternates during the first year I am a member, paying in most cases special members' prices. My membership is cancelable any time after I buy these four books. A shipping charge is added to all shipments. **65**

MR. / MRS. / MISS .....
(Please print plainly)

Address..... *Apt*.....

City & State..... *Zip*.....

FACTS ABOUT MEMBERSHIP

• You will receive the Book-of-the-Month Club *News*, a literary magazine published by the Club fifteen times a year. The *News* describes the coming Selection and scores of Alternates, and will be sent to you approximately every three and a half weeks.
• If you wish to purchase the Selection, do nothing and it will be shipped to you automatically.
• If you do not want the Selection – or you would like one of the Alternates or no book at all – simply indicate your decision on the reply form always enclosed with the *News* and mail it so we receive it by the date specified.
• If, because of late mail delivery of the *News*, you should receive a Selection without having had at least 10 days to decide whether you want it, that Selection may be returned at Club expense.
• If you continue after your trial membership, you will earn at least one Book-Dividend® Credit for every Selection or Alternate you buy. These Credits enable you to acquire for your library a wide variety of books, called Book-Dividends, at astonishing savings – *at least 70% of publishers' list prices.* Under this unique system, more than 100 Book-Dividends will be available to choose from every year.
• All books distributed by the Book-of-the-Month Club are identical to the publishers' editions in content, format, size and quality.

Problems can be opportunities in disguise.

The twin dilemmas of energy and the environment may, in the long run, provide man with his best chance to reassess some old assumptions and reorder his priorities for a safer, more pleasant life.

Even before these problems reached crisis proportions, Chiyoda—Japan's leading engineering firm—was developing technology that will help ensure both environmental integrity and efficient energy production.

For example, to minimize air pollution and alleviate the shortage of low-sulfur fuels, Chiyoda has built about half the fuel oil desulfurizers in Japan. We've developed Chiyoda Fume Cleaner to collect the most harmful fume, consisting of particles of less than 3.0 microns. And we've introduced the Chiyoda THOROUGHBRED series of pollution control systems: new processes for flue gas desulfurization, the simultaneous removal of sulfur dioxide and nitrogen oxides, and waste water treatment.

These days "technology assessment" has become a fashionable expression. In our fields, we've been practicing it for years.

RESPONSIBLE, GLOBAL ENGINEERING

CHIYODA CHEMICAL ENGINEERING & CONSTRUCTION CO.,LTD.

Yokohama, Japan. Telex: CHIYO J47726

Overseas Offices: U.S.A.: Chiyoda International Corporation 1300 Park Place Building, 1200 6th Ave., Seattle, Wash. 98101 **England: Chiyoda Representative Office** 16th Floor, St. Alphage House, Fore Street, London E.C. 2 **Singapore: Chiyoda Singapore (Pte.) Ltd.** 8-E, Chia Ping Road, Jurong Town, Singapore 22 **Iran: Chiyoda Iran Co. (Private)** No. 56 Takhte Jamshid Avenue, Teheran

Established 1845 **SCIENTIFIC AMERICAN** August 1975 Volume 233 Number 2

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $15 PER YEAR, U.S., POSSESSIONS AND CANADA; $18 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O PUBLICATIONS DISTRIBUTION SERVICE, ST. KATELIJNEVEST 14, B2000 ANTWERP, BELGIË.

THE COVER

The painting on the cover is a representation of an experimental arrangement in a planetarium for testing a bird's ability to orient itself by the stars. In April and May the indigo bunting normally migrates from its wintering grounds in the Bahamas, southern Mexico and Central America to its breeding grounds in the eastern U.S. A bird held in captivity will display intense nocturnal activity during its migratory period. When a bunting in a migratory condition is placed in a circular test cage, shown here with its front portion cut away, and exposed to the pattern of stars that appear in the northern sector of the sky, the bird orients its hopping activity to the north, its normal migratory direction. The hops are recorded on the white interior of the test cage because the bird stands on an inked pad (see "The Stellar-Orientation System of a Migratory Bird," page 102).

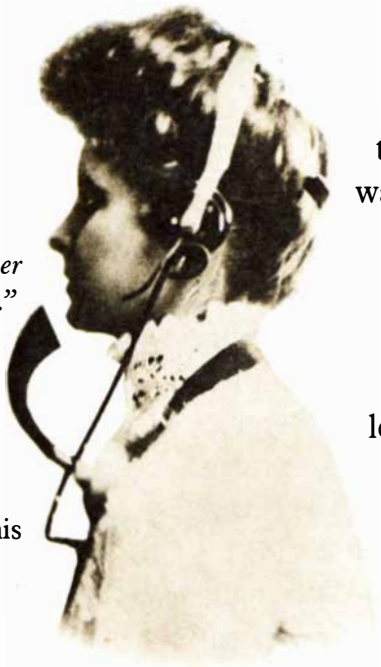
THE ILLUSTRATIONS

Cover painting by Richard Ellis

Page	Source
17–25	Total Communications Industries Ltd.
27	Leiden Observatory
28	Aerophoto Eelde
29	George V. Kelvin
30	George V. Kelvin (*top*), Leiden Observatory (*bottom*)
31–33	George V. Kelvin
34	Leiden Observatory
35	George V. Kelvin
37	H. Kutzner, Technische Hochschule, Darmstadt (*top*); Ginny Fonte, University of Colorado (*bottom*)
38	Norbert Pfennig, University of Göttingen
39–41	Bunji Tagawa
42	Julius Adler, University of Wisconsin (*top*); Bunji Tagawa (*bottom*)
43	Howard C. Berg, University of Colorado
44	Bunji Tagawa
51	Howard S. Friedman
52	David T. Horsley
53–54	Lorelle M. Raboni
55	P. C. Hillier
56–60	Lorelle M. Raboni
63	Ellen Rixford
66–67	Dan Todd
68	Ben Rose (*top*); Jacob Beck, University of Oregon (*bottom*)
69	Jacob Beck, University of Oregon
72–75	Dan Todd
78	Woods Hole Oceanographic Institution
80–82	Alan D. Iselin
83	Woods Hole Oceanographic Institution (*left*), Alan D. Iselin (*right*)
84	Alan D. Iselin
85	Alan D. Iselin (*top*), Woods Hole Oceanographic Institution (*bottom*)
86	Woods Hole Oceanographic Institution
87–89	Alan D. Iselin
92–93	Melvin L. Fowler, University of Wisconsin at Milwaukee
94–99	Tom Prentiss
100–101	Illinois State Museum
102–103	Ralph Morse
104	Adolph E. Brotman
105	Adolph E. Brotman (*top*), Ralph Morse (*bottom*)
106–111	Adolph E. Brotman
112–115	Ilil Arbel
116–119	Jerome Kuhl

One of a series of reports on the first hundred years of the telephone.

How there came to be only one telephone company in town.



"Number please."

"In many cities of the United States, and in rural communities as well, there are dual and competing telephone systems, doing both local and long-distance business...Patrons of these telephone systems are put to endless annoyance and increased expense. In order to reach all the people using telephones, the telephone patron finds he must install two telephones in his house and office...Double systems of cables, wires and conduits burden the streets and highways."

—Report of the House of Representatives Committee on Interstate and Foreign Commerce, 67th Congress (1921)

When Alexander Graham Bell's telephone patents expired in 1893 and 1894, new telephone companies sprang up almost overnight. The accepted way of organizing communications was to have the "dual and competing telephone systems" cited in the Congressional report.

"Call us. We're on the Bell," was a frequent invitation in those days, to friends or customers. Central, the voice of "Number, please?", spent a lot of time explaining to customers that the number wanted was on the town's other telephone system. And each month there were two telephone bills to pay.

A solution to the problem had been worked out long before by John Stuart Mill. In 1847 Mill had studied the situation of two other new industries that supplied water and gas through pipes to the homes and businesses of London:

"It is obvious, for example, how great an economy of labour would be obtained if London were supplied by a single gas or water company instead of the existing plurality. While there are even as many as two, this implies double establishments of all sorts, when only one, with a small increase, could probably perform the whole operation equally well; double sets of machinery and works, when the whole of the gas and water required could generally be produced by one set only; even double sets of pipes, if the companies did not prevent this needless expense by agreeing upon a division of the territory. Were there only one establishment, it could make lower charges, consistently with obtaining the rate of profit now realized."

Such a consolidation, Mill saw, was clearly in the public interest. The concept of a "public utility" was reinforced.

John Stuart Mill

When Edison's electric light superseded illuminating gas, the parallel was obvious. It was not quite so obvious for the telephone.

It was not hard to see that the public benefited from having water piped into homes. But while some viewed the telephone companies as providing a similar vital service, others regarded them as being more akin to manufacturers selling ingenious machines in the luxury class. When only a few people had telephones, one observer called them "electric toys." Should Bell's invention be compared with Edison's new electric light, or was it more like his phonograph? As the proportion of homes and businesses with telephones grew, the usefulness of the telephone increased greatly.

Then there was the matter of geographic area served. An exclusive franchise for a specified area is a natural corollary of Mill's concept of a public utility. And exclusiveness was a troublesome subject.

When two or more rivals supply a similar service, competition keeps each up to the mark, or else some eventually lose customers and go out of business. If in the public interest, government removes that rivalry by granting exclusive franchises, then government must provide the mechanisms for preventing arbitrary or excessive charges or unreasonable or discriminatory regulations.

The doctrine of public regulation of privately owned resources has its roots in Roman law and the tenet of *justum pretium* —"just price." English common law provided a rationale for regulation. In an essay on rates for wharf services, Sir Matthew Hale, Lord Chief Justice of England, established in 1670 the criterion that private industries "affected with a public interest" may be regulated by the public:

> "If the King or subject have a public wharf unto which all persons that come to that port must come and unload their goods...because they are the only wharfs licensed by the King...or because there is no other wharf in that port...there cannot be taken arbitrary and excessive duties...but the duties must be reasonable and moderate....For now the wharf and crane and other conveniences are affected with a public interest."

Various municipal boards did undertake to control the quality of service provided by water, gas and electric companies, usually through periodic reviews of franchises granted. It was no easy task. For quality of service leads quickly to questions of cost: good service for the price charged;

equal prices for all customers for services of a similar nature, so that no one is discriminated against; adequate service capacity so that anyone able to pay for the service can have it.

Local officials had their hands full regulating the three industries already mentioned (water, gas, electricity). They were not eager to take on the responsibility of regulating the telephone business. And so redundant companies continued to exist in many towns. Confusion multiplied geometrically as the companies strung long distance lines to connect various cities.

Theodore N. Vail

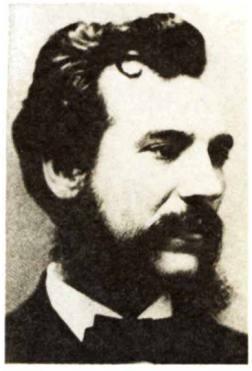
Soon after he was elected A.T.&T. President in 1907, Vail enunciated the goal: "One policy, one system, universal service." He saw that the future of the business depended on having one unified telephone service for the entire nation—a service that every family and business could enjoy. That meant ending duplicate telephone companies, replacing them with exclusive telephone franchises. In other words, Vail understood that it was not enough for the nation to have telephone companies. What was needed—and what he sought to create—was a *telephone system.* Vail saw, too, that the very "exclusivity" of the franchises invited—indeed, demanded—regulation by officials elected or supported by the public to protect the public interest.

Vail thus agreed with the efforts of Gov. Charles Evans Hughes of New York and Senator Robert M. LaFollette of Wisconsin, who were working to persuade state legislatures to try a new approach to regulation through state utility commissions—responsive to the public at the state level—as best serving the public interest.

The state commissions, supported by public desire for efficient regulation, worked. Most public utilities came to be regulated on a statewide basis, and a framework of efficient regulation was set.

Vail recognized, however, that national regulation also was a necessary complement to state regulation, particularly since one company—A.T.&T.—was chiefly responsible for interconnecting the individual telephone companies into a telephone system.

As noted earlier, regulation of telephone companies already had begun to develop at the state level. But on the national level—for telephone lines and services crossing state boundaries—there was no federal counterpart to the state regulatory commissions, although telegraph companies had been regulated to some extent by the Postmaster General and the Interstate Commerce Commission under statutes dating back to the 19th century.



Alexander Graham Bell

The first effort at comprehensive federal regulation came in the Mann-Elkins Act of 1910, amending the Interstate Commerce Act. Telephone, telegraph and cable companies were declared to be common carriers subject to ICC regulation.

Federal regulation took a new turn in 1934 with the passage by Congress of the Communications Act, which established the Federal Communications Commission. The intent of Congress—as it had been the intent of both Bell and Vail—is outlined in Section I of the Communications Act:

"For the purpose of regulating interstate and foreign communications by wire and radio so as to make available, so far as possible, to all the people of the United States a rapid, efficient, nationwide and worldwide wire and radio communication service with adequate facilities at reasonable charges...."

Almost alone among the nations of the world, then, this country entrusted the development and operation of its communications resources to private enterprise. It endowed companies with the rights and responsibilities of common carriers, each solely privileged to purvey its services within its territory but all in turn strictly accountable through state and national regulation to the public they serve.

Has it worked?

In 1968, President Johnson's Task Force on Communications Policy concluded, "It can be truly said that the United States has the finest telephone system in the world."

That kind of service didn't just happen. It was planned that way, right from the start.

One Bell System. It works.

Bell System

LETTERS

Sirs:

Stillman Drake's conjecture about Galileo's timekeeping [SCIENTIFIC AMERICAN, June], namely that, as a musician, he could maintain an accurately constant tempo, brings to mind a similar stratagem by which Benjamin Franklin measured relative times without a watch or a clock. (Letter to Sir John Pringle, May 10, 1768.)

In 1767, while on a canalboat in Holland, Franklin was puzzled by the boatman's explanation for the low speed of the boat, namely "that it had been a dry season, and the water in the canal was low[, not] so low as that the boat touched the muddy bottom;... but so low as to make it harder for the horse to draw the boat." Confirmation by other boatmen convinced him that the resistance of a boat does indeed increase when the boat is in shallow water.

In seeking to explain the effect of water depth on water resistance, Franklin reasoned that "the boat in proceeding along the canal, must in every boat's length of her course, move out of her way a body of water, equal in bulk to the room her bottom took up in the water; that the water so moved, must pass on each side of her and under her bottom to get behind her; that if the passage under her bottom was straitened by the shallows, more of that water must pass by her sides, and with a swifter motion, which would retard her, as moving the contrary way."

In order to study the effect experimentally Franklin set up a long, narrow towing tank in which he could adjust the depth of the water by adjusting the level of a false bottom. Along this model canal he towed "a little boat in form of a lighter or boat of burthen" by means of "a long silk thread [attached] to its bow," with "the other end passed over a well made brass pully...; and a shilling was the weight. Then placing the boat at one end of the trough, the weight would draw it through the water to the other." Franklin timed the passage of the model canalboat in three depths of water. Averaging his times for eight passages at each depth, he found that the time difference "between the deepest and shallowest... appears to be somewhat more than one fifth."

How did Franklin measure time? "I counted as fast as I could count to ten," he wrote, "keeping an account of the number of tens on my fingers."

The effect that Franklin studied is closely related to a wind-tunnel effect that we must take into account in our aerodynamic studies at the Langley Research Center. The wind-tunnel walls constrict the flow of air about a model so that the airspeeds in the immediate neighborhood of the model are increased.

SAMUEL KATZOFF

Langley Research Center
Hampton, Va.

Sirs:

Your readers are entitled to a correct account of G. D. Birkhoff's theory of gravitation. Contrary to what Clifford M. Will says in "Gravitation Theory" [SCIENTIFIC AMERICAN, November, 1974], the Birkhoff theory did not demand "that sound waves travel at the speed of light, in violent disagreement with fact." It had nothing to do with sound waves.

Like most relativistic theories, it was intended to account for the three "crucial effects": the advance in the perihelion of the planet Mercury, the bending of light around the sun and the red shift in spectral lines from distant stars. Like Einstein's original general theory of relativity, it was speculative. As everyone should know, although Einstein's special theory of relativity in "flat" space-time is a universally accepted cornerstone of 20th-century physics, his general theory is only one of many models proposed to explain the three crucial effects. As Will says, even among "metric" theories there are many others.

After a quest of two decades, G. D. Birkhoff found a simple model that easily predicted the first two effects. The red shift was another matter. Birkhoff proposed *two* explanations of it. The first is based on a "perfect fluid," which he had proposed in 1926 as an alternative to Schrödinger's equations for predicting the observed spectrum of hydrogen. Although his "perfect fluid" had the advantage of conforming to the Lorentz group (which Schrödinger's equations do not), its status is analogous to that of Einstein's speculations about a "unified field theory." Birkhoff resurrected his "perfect fluid" in his first attempt to explain the red shift. He gave a popular exposition of it and other theories of gravitation in the January and February 1944 issues of *The Scientific Monthly*.

In your pages Will was apparently summarizing an interpretation proposed by others, according to which, in collapsed dense stars, the Birkhoff "perfect fluid" would predict that "pressure" and gravitational waves would propagate with the speed of light. Whether or not my father would have agreed with this interpretation, God only knows.

In any case, shortly before his death in November, 1944, G. D. Birkhoff proposed a second explanation of the red shift that did not involve his perfect fluid. Using Einstein's equation $h\nu = E$, he attributed the red shift to "the energy change of the photon as it travels from the emitting body to the earth." In this final model, although gravitational waves travel with the speed of light, sound waves certainly do not.

GARRETT BIRKHOFF

Cambridge, Mass.

Sirs:

I greatly enjoyed Professor Wilson's June article on slavery in ants, in which he describes how certain species of ants steal the pupae of other ants. If ants could speak, perhaps the victimized species would sing: "Let my pupae go!"

JAMES METCALFE

New York

Scientific American, August, 1975; Vol. 233, No. 2. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the author's risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Harry T. Morris, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 30 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on subscription is shown at upper right-hand corner of label.)

NAME

NEW ADDRESS

OLD ADDRESS

Just two of the great performers whose latest recordings are now offered on stereo records, cartridges and cassettes

THE CLASSICAL DIVISION OF COLUMBIA MUSIC CLUB INVITES YOU TO TAKE

any 8 for only $1.00

plus shipping and handling

if you join now and agree to buy only 8 selections (at regular Club prices) in the coming two years

YES, Leonard Bernstein and Sir Georg Solti are but two of the great conductors whose thrilling performances are now offered on these pages. You'll also find superb recordings by Ashkenazy, Boulez, Horowitz, Mehta, Ormandy, Serkin, Stern (to name but a few) . . . in widely-acclaimed performances of the greatest symphonies, concertos, orchestral masterpieces of all time! And as a new member, you may take **ANY 8** of these selections — in your choice of stereo records, cartridges or cassettes — **all for only $1.00.**

Your only membership obligation is to buy just eight selections in the coming two years, at the regular Club prices (which currently are: $6.98 for records; $7.98 for cartridges and cassettes, plus shipping and handling. Multi-unit sets and Double Selections may be somewhat higher).

Think of it . . . you need buy as few as four records or tapes a year . . . and you won't have any difficulty in selecting them, because you'll receive our Classical Bulletin regularly every four weeks (13 times a year). And we'll also send you the regular Club Magazine, which contains additional classical recordings, as well as hundreds of selections from every field of music. In addition, up to six times a year we may send some special issue of the Bulletin, offering extra-special savings.

This Bulletin will tell you about many of the newest classical releases and we'll even select one as "The Selection of the Month" — but there is no obligation to accept it . . . order only those you want *when you want them!*

A special response card will be enclosed with each Bulletin . . . and you may use it to order or reject any selections offered simply by mailing it by the date specified. And if you want *only* The Selection of the Month, you need do nothing — it will be shipped to you automatically. What's more, you'll have at least ten full days in which to make your decision — if you ever have less time than that, just return The Selection of the Month at our expense, for full credit.

You may cancel membership at any time after you've purchased eight selections — or continue under our special bonus plan, which offers you at least 33% savings on all future purchases. Here, indeed, is the most convenient way possible of acquiring the classical library you want — *at the greatest savings possible!* So do act now.

249730-249731. Tchaikovsky: The Nutcracker (complete) — Rozhdestvensky, Bolshoi Orch. (2-record set counts as 2 — Columbia/Melodyia)

245381. Tchaikovsky: Symphony No. 5 — Ormandy, Phila. Orch. (Columbia)

138909-138900. Tchaikovsky: The Sleeping Beauty *and* **Swan Lake Ballet Suites** — Hollingsworth, Sinfonia London (Columbia Treas.)*

235184-235185. Richard Tucker and Robert Merrill at Carnegie Hall — duets and solos by Verdi, Bizet, Mozart, etc. (2-record set counts as 2 — London)

195024-195025. Verdi: Requiem — Bernstein, London Symphony Orch., Chorus (2-record set counts as 2 — Columbia)*

246645. Music From Walt Disney's "Fantasia" — *Night on Bald Mountain* (Stokowski, London Sym.); *Sorcerer's Apprentice* (Hermann, London Phil.); many others (London)*

227561. Vivaldi: The Four Seasons — violinist Pinchas Zuckerman and the English Chamber Orch. (Columbia)

233684. Wagner: Tristan und Isolde (*Prelude and Love-Death,* etc.)—Boulez, N.Y. Phil. (Columbia)

219881. John Williams plays **Great Guitar Hits** by Sor, Albeniz, Bach, Vivaldi, etc. (Columbia)

Note: selections with two numbers are 2-record sets or double-length tapes. Each of these "double selections" counts as 2 -- write in both numbers.

Columbia House
1400 North Fruitridge Avenue
Terre Haute Indiana 47808

All applications subject to review and Columbia House reserves the right to reject any application

Columbia Music Club/CLASSICAL DIVISION
Terre Haute, Indiana 47808

Please accept my membership application in the Club under the terms outlined in this advertisement. Send me the 8 classical selections listed below for only $1.00, plus shipping and handling. I agree to buy 8 selections (at regular Club prices) during the coming two years — and may cancel membership at any time after doing so. **I am interested in the following type of recordings (CHECK ONE ONLY):**

☐ **RECORDS** ☐ **CARTRIDGES** ☐ **CASSETTES**

2ML/GF

Write in numbers of 8 selections		

☐ **Mr.**
☐ **Mrs.**
☐ **Miss** .....
(Please Print) **Last Name** **First Name** **initial**

Address ..... **Apt. No.** .....

City .....

State ..... **Zip** .....

Do You Have A Telephone? (Check one) ☐ **YES** . . ☐ **NO**

APO, FPO, Alaska, Hawaii: write for special offer 556/F75

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1925: "Einstein's prediction of the shift of spectrum lines toward the red when they are produced in a strong gravitational field has been verified by observations of the faint companion star of Sirius. This star has only one ten-thousandth the brightness of Sirius but 40 percent of its mass. On this basis Eddington has calculated that its diameter should be only one thirty-fifth that of the sun, but this means the star's density should be 50,000 times that of water. Until last year this seemed absurd, but Eddington has shown that in the inside of a star the fragments of atoms should be capable of being driven very close together, so that even this enormous density is quite believable. Because the star is so small and massive, the Einstein shift (which is proportional to the mass divided by the radius) should correspond to a velocity of 20 kilometers per second and thus should be measurable. The spectrum has now been photographed several times at Mount Wilson. The observed shift comes out to more than 30 kilometers per second, corresponding to a shift of the lines in the direction Einstein predicted and to an even greater amount than Eddington calculated. It appears, therefore, that Einstein's prediction is verified, and that the companion of Sirius is even smaller than Eddington's calculations made it."

"A hornless loudspeaker of new design has been invented by two engineers of the General Electric Company, C. W. Rice and E. W. Kellogg. The varying currents from the radio set are passed through a coil, which is thus caused to vibrate. The coil is attached to a diaphragm, a paper cone about six inches in diameter. An important feature of the loudspeaker is the baffle board, which serves as the front of the cabinet. The use of the baffle makes it possible to dispense with horns without sacrificing the radiation of the deeper tones."

"The Sicilian corner on sulfur, managed by an English syndicate, has been entirely broken through the success of chemical engineering in America. It was known that 1,000 feet below the surface in Louisiana and Texas deposits of sulfur were to be found. Several attempts to reach them had resulted only in a loss of life and investment. The substance is now extracted by forcing water and steam down to the deposit, where the sulfur melts and flows to the surface."

"The time has come to put an end to the abominable nuisance caused by oil-burning ships discharging residue from their tanks on the high seas and within the harbors and ports of the world. This filthy mess is defiling many of the most attractive bathing resorts; it is endangering the shipping in our harbors because of the liability of the floating oil catching fire, and it is causing widespread destruction of the bird life of the seas. The floating refuse gums up the feathers and the wings of sea fowl to such an extent that they are unable to lift themselves from the water, cannot move about in search of food, and simply drift as starving derelicts until they are cast up by the thousand upon the beach."

SCIENTIFIC AMERICAN

AUGUST, 1875: "No intelligent student of Science can help being struck with the violations of the first principles of mechanical philosophy in the current conceptions of ether and dense matter. If we assume the ether to be continuous, then solid atoms moving through it must displace their own bulk and hence lose force of motion. According to the laws of fluid resistance, all bodies moving within the ether would be resisted according to the squares of their velocities, yet we observe the unresisted motions of the heavenly bodies. The matter is not mended by supposing the ether to be particled and ascribing definite intervals between the particles. 'The intensity of the light depends on the distance to which the ether particles move to and fro,' Tyndall writes. He does not say whether the ether particles strike against one another or not. If they do, they have only a certain amount of space to swing in; if they do not, we are thrown back upon the generally repudiated theory of action at a distance."

"Immense boulders, called 'lost rock,' which have no resemblance to any mass of rock in their vicinity, are scattered over the northern part of the United States. Similarly, there are heaps of sand, gravel and cobble stones which form many of our ridges, knolls and hills and which are totally unlike any fixed rock near them. To explain the transportation of these wanderers from their homes various theories have been advanced, such as the effects of floods, or of powerful mud currents or of gas explosions that hurled rocks in all directions. These and many others, however, fail to satisfy the observed conditions. To Louis Agassiz belongs the credit of first attributing all these phenomena to glacial action. Agassiz conceived of a sheet of ice and snow extended enough to cover a continent. Having noticed that markings below glaciers in the Alps were the same as those found beneath the ice mass, he compared these with similar appearances in northern Europe and Asia, which were generally attributed to the Great Flood, and made the bold generalization that all were due to the very same cause, and that one vast sheet of ice must have covered all the northern regions of the globe."

"Mr. Isaac M. Singer, the sewing machine inventor, recently died in England. In early life Mr. Singer worked at the machinist's trade, but on the appearance of the sewing machine he at once turned his attention in that direction. His enterprise did not go long unrewarded: his corporation became one of that combination that practically controlled the sewing machine trade of the country."

"In the United States little is known concerning industrial progress in Russia, but from the death of Peter the Great in 1725 Russian manufacturers have steadily pushed onward. The manufacture of iron holds a very important position. Sheet iron is produced to a great degree in large private establishments, but steel-making is yet in its infancy, the metal being made almost entirely by one or two government foundries. It is applied chiefly to the manufacture of cannon. In some districts the manufacture of cutlery and hardware forms the sole occupation of the entire male population, and Russian iron is now largely employed in the making of cut nails. Samovars are also a leading article of the Russian metal trade; their material is copper, which is almost exclusively used among the well-to-do classes for cooking utensils. It may be said that Russian manufactured goods have for the most part attained a high degree of excellence, but many of them are enormously dear. The interests of the mass of Russian people who consume are thus sacrificed to increase the wealth of the comparatively small class who manufacture."

THE AUTHORS

KERR L. WHITE ("International Comparisons of Medical Care") is professor of health-care organization at Johns Hopkins University. A graduate of McGill University, he was trained in medicine and also did graduate work in industrial management and epidemiology at Yale University and the University of London. Before joining the faculty of Johns Hopkins he taught medicine at the University of North Carolina and the University of Vermont. White served as chairman of the study of medical-care utilization sponsored by the World Health Organization; the results of that study are summarized in his article. He wishes to acknowledge the assistance of Renate Wilson, Jane H. Murnaghan, Barbara Vaeth and Donald O. Anderson in the preparation of the article.

RICHARD G. STROM, GEORGE K. MILEY and JAN OORT ("Giant Radio Galaxies") are astronomers at the Leiden Observatory. Strom, a native of New York, did his undergraduate work in physics at Tufts University and went on to earn his master's degree in low-temperature physics from Dalhousie University. Turning to radio astronomy, he then went to the University of Manchester's Nuffield Radio Astronomy Laboratories at Jodrell Bank, where he received his Ph.D. in 1972. A member of the research staff at Leiden since 1970, he will take up a new position this fall as astronomer with the Netherlands Foundation for Radio Astronomy. Miley, who was born in Ireland and majored in physics as an undergraduate at University College, Dublin, also began his astronomical career at Jodrell Bank, acquiring his Ph.D. from Manchester in 1968. He worked for two years as a research associate of the U.S. National Radio Astronomy Observatory before joining the Leiden Observatory in 1970. Oort was born in 1900 "in the small Frisian city of Franeker, where Eyse Eysinga built his famous planetarium." He began his association with the Leiden Observatory in 1924, soon after completing his studies at the University of Groningen; since his retirement from the observatory staff in 1970 he has continued his research there. His major work has been on the structure and dynamics of galaxies and on the origin of comets.

HOWARD C. BERG ("How Bacteria Swim") is professor of biology at the University of Colorado. As an undergraduate at the California Institute of Technology he studied physics and chemistry. After spending a year as a Fulbright scholar at the Carlsberg Laboratory in Copenhagen he entered the Harvard Medical School, but he withdrew two years later to "cross the Charles" and complete a thesis on the atomic hydrogen maser for his Harvard Ph.D. in experimental physics. He spent three years as a junior fellow in the Harvard Society of Fellows before joining the Harvard faculty in 1966. He made his move to Colorado in 1970. In addition to his primary research interest in the motile behavior of bacteria, he reports, "I am also doing some work on the chemical modification of cell membranes."

BRYAN CLARKE ("The Causes of Biological Diversity") is professor of genetics at the University of Nottingham, where he has spent the past four years setting up a new genetics department. He went to Nottingham from the University of Edinburgh, where he had taught zoology for 12 years after receiving his D.Phil. from the University of Oxford. He writes: "I am obsessively interested in genetics and ecology. From time to time I am also interested in gardening, woodworking, traveling, photography, flying, skiing, writing and arguing."

JACOB BECK ("The Perception of Surface Color") is professor of psychology at the University of Oregon. A graduate of Yeshiva University, Beck went on to Cornell University, where he obtained his Ph.D. in psychology in 1958. He has taught at the University of Pennsylvania and Harvard University. During 1972–1973 he was a visiting member of the computer-science center of the University of Maryland, where he worked on problems of computerized pattern recognition. The following year he took a leave of absence to serve as program director for psychobiology at the National Science Foundation. In addition to his studies on color perception he has conducted research "on the visual perception of form and pattern and on the auditory perception of sound and pitch."

J. R. HEIRTZLER and W. B. BRYAN ("The Floor of the Mid-Atlantic Rift") are at the Woods Hole Oceanographic Institution. Heirtzler, who serves as chairman of the department of geology and geophysics at Woods Hole, is a graduate of Louisiana State University. His doctorate is from New York University. Before taking up his present position in 1969, he was director of the Hudson Laboratories of Columbia University. He has also taught at the American University of Beirut, worked as a physicist for the General Dynamics Corporation and done research at Columbia's Lamont-Doherty Geological Observatory. A specialist in marine geomagnetism and its relation to sea-floor spreading, he was chief scientist on the Project FAMOUS exploration of the Mid-Atlantic Rift. Bryan, a petrologist and a member of the Project FAMOUS diving team, has his B.A. from Dartmouth College and Ph.D. from the University of Wisconsin. He developed an interest in volcanic activity around the ocean basins while working in the Aleutians in 1954 as a field geologist for the U.S. Geological Survey. In 1962 he took a position on the staff of the University of Queensland "in order to have further opportunities to gain familiarity with geologic problems in the southwest Pacific area." He left Queensland in 1967 to join the Carnegie Institution Geophysical Laboratory in Washington, D.C., and has been at Woods Hole since 1970.

MELVIN L. FOWLER ("A Pre-Columbian Urban Center on the Mississippi") is professor of anthropology at the University of Wisconsin at Milwaukee. A Meso-American by birth (Nebraskan), he was educated at Purdue University and the University of Chicago, receiving his Ph.D. from Chicago in 1949. Before moving to Wisconsin in 1964 he was affiliated with the Illinois State Museum and Southern Illinois University. His chief research interest lies in "the archaeological study of the temple-town community in different environments as a prelude to urban development"; his territory is "Meso-America, in particular the central Mississippi River valley."

STEPHEN T. EMLEN ("The Stellar-Orientation System of a Migratory Bird") is associate professor of animal behavior at Cornell University. After graduating from Swarthmore College in 1962, he went to the University of Michigan, where he obtained his Ph.D. in zoology in 1966. He writes: "I now split my time between two quite different research areas within animal behavior: migration, orientation and navigation on the one hand and social behavior on the other.... This summer my wife (who is also a researcher on animal behavior) and I will be traveling to the Arctic to collect eggs and baby chicks of sandpipers to bring back to Cornell for in-depth studies of the ontogenetic development of orientational abilities."

The SR-50A. The SR-51A.

Texas Instruments is steeped in calculator technology from start to finish. We make all critical parts, and control quality every step of the way. This is the key to the exceptional quality and value of TI's professional calculators.

Capability. Quality. Value.

The technological achievement under the keyboard is still the reason TI's professional calculators offer so much quality and math power for the money.

Engineer . . . Scientist . . . Businessman . . . Geologist . . . Chemist . . . Statistician . . . Student . . . whatever your field, if you're doing more than basic mathematics, consider an SR-50A or SR-51A from Texas Instruments.*

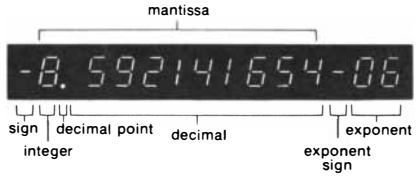
SR-51A: simple arithmetic to complex statistics.
Sheer math power. Log and trig and hyperbolics and functions of x. The SR-51A has these and also statistical functions. Like mean, variance and standard deviation. Factorials, permutations, slope and intercept. Trend line analysis. And there's a random number generator. Plus 20 preprogrammed conversions and inverses. Check this list for a closer look at the real math power you can get in both the SR-51A and the SR-50A:

FUNCTION	SR-51A	SR-50A
Log, lnx	yes	yes
Trig (sin, cos, tan INV)	yes	yes
Hyperbolic (sinh, cosh, tanh, INV)	yes	yes
Degree-radian conversion	yes	yes
Deg/rad mode selection switch	yes	yes
Decimal degrees to deg.min.sec.	yes	no
Polar-rectangular conversion	yes	no
y^x	yes	yes
e^x	yes	yes
10^x	yes	no
x^2	yes	yes
$\sqrt{x}$	yes	yes
$\sqrt[y]{}$	yes	yes
1/x	yes	yes
x!	yes	yes
Exchange x with y	yes	yes
Exchange x with memory	yes	no
% and Δ %	yes	no
Mean, variance and standard deviation	yes	no
Linear regression	yes	no
Trend line analysis	yes	no
Slope and intercept	yes	no
Store and sum to memory	yes	yes
Recall from memory	yes	yes
Product to memory	yes	no
Random number generator	yes	no
Automatic permutation	yes	no
Preprogrammed conversions	20	1
Digits accuracy	13	13
Algebraic notation (sum of products)	yes	yes
Memories	3	1
Fixed decimal option	yes	no
Keys	40	40
Second function key	yes	no
Constant mode operation	yes	no

Performance, accuracy and efficiency.
Both the SR-50A and SR-51A deliver answers you can trust. Quickly and efficiently. To problems ranging from simple arithmetic to highly complex calculations. You don't have to learn special entry methods or difficult-to-master key sequences. There's a better way – TI's algebraic entry system lets you key your problem just the way you would say it. Naturally. You don't worry about losing data in stacks, or keeping track of what is in each stack, or remembering if the stack is full. The way you learned math is the way it's done. On both the SR-50A and SR-51A – you can command tremendous math power with confidence – from the beginning. Power and accuracy you can really put to work.

Answers are calculated to 13 significant digits, rounded off and displayed to 10. And for maximum accuracy, all 13 are held inside for subsequent calculations.

Scientific notation is automatic when you need it. For numbers as large as $\pm 9.999999999 \times 10^{99}$. Or as small as $\pm 1. \times 10^{-99}$.



Quality craftsmanship.
Quality – it's built in right from the start. Texas Instruments designs and manufactures every critical component. From high-purity silicon semiconductor materials to integrated circuits to light-emitting-diode displays to circuit boards to keyboards. So, we design-in and control quality – not just monitor it – at every level: Materials. Components. The complete system.

To assure you reliable performance, every calculator is subjected to severe environmental and reliability testing prior to release to production. In production, every one is thoroughly tested, then "burned-in", then thoroughly tested again. If there's any problem, we want to find it before it gets to you.

Inside, steel machine screws anchor all important structural elements – plastic welds and glue fastenings aren't good enough. A double-tough Mylar** barrier keeps dust and moisture from getting under the keyboard. The case is high-strength, injection-molded plastic designed to take a beating. It's a quality calculator. And you know it as soon as you get your hands on one. The heft and solid feel tells you it's a fine-quality instrument even before you press a key.

The SR-50A and SR-51A are human engineered, too, for maximum comfort and efficiency. For a hand or a desktop. Keys have positive-action, tactile feedback. And the big, bright displays are easy to read at your desk or on the go. Slim. Compact. Light. In your briefcase or on your belt, you'll hardly notice just 8.3 ounces.

New, low SR-50A and SR-51A prices.
Technological leadership and quality craftsmanship are why Texas Instruments can offer so much value at low prices. And now, with new price tags, the SR-50A and SR-51A are better values than ever before: $99.95 for the SR-50A. $149.95 for the SR-51A.

SR-51A Preprogrammed Conversions

FROM	TO
mils	microns
inches	centimeters
feet	meters
yards	meters
miles	kilometers
miles	nautical miles
acres	square feet
fluid ounces	cubic centimeters
fluid ounces	liters
gallons	liters
ounces	grams
pounds	kilograms
short ton	metric ton
BTU	calories, gram
degrees	gradients
degrees	radians
°Fahrenheit	°Celsius
deg.min.sec.	decimal degrees
polar	rectangular
voltage ratio	decibels

See them at your nearest TI calculator retailer. Or, send for our new fact-filled color brochure. It details the outstanding capability of both the SR-50A and SR-51A with full feature descriptions, sample problems, entry-method considerations and more. Write, Texas Instruments, M/S358, P.O. Box 22013, Dallas, Texas 75222

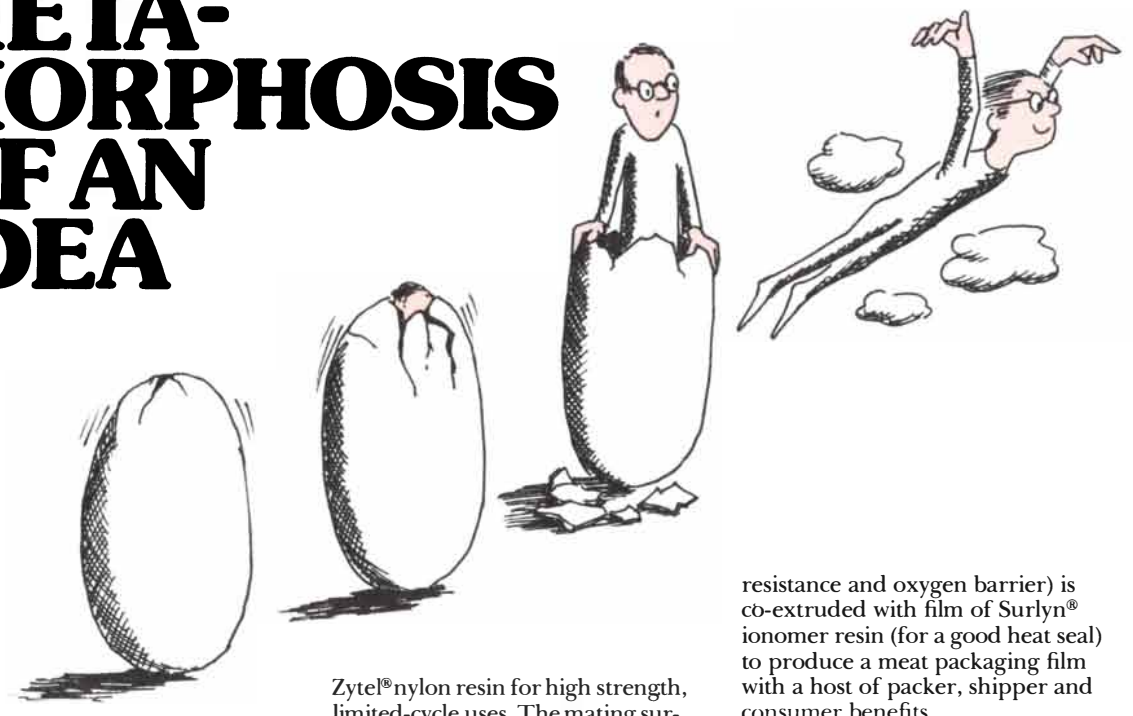
TEXAS INSTRUMENTS
INCORPORATED

* The SR-50A and SR-51A are our popular SR-50 and SR-51 in handsome new case designs

** Trademark of DuPont

82073

META-MORPHOSIS OF AN IDEA



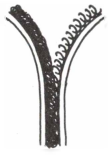
Ideas. Where in the world do they come from? The process is as mysterious as any in nature. We only know that they occur, sometimes unasked, often haphazardly. Some of the best ideas come as a perceptive statement of the problem whether or not the situation was ever before thought of as a problem.

Sticky problem

Ideas can come by studying natural phenomena. Consider Swiss engineer George de Mestral. He set out to develop a low-cost fastener that would overcome the problems of lost buttons, broken hooks and sticking zippers.

He used a practice dear to the scientific heart—drawing an analogy from Mother Nature. He remembered how burdock burrs clung to his pants. So he put the burrs under the microscope and saw that they consisted of tiny hooks that grabbed onto any thread or hair. Eureka!

Mr. de Mestral's original idea and Du Pont technology were combined to give the world VELCRO® brand fasteners. You've probably seen them. One surface is covered with stiff, tiny, woven hooks. All lined up in straight, narrow rows and—unlike nature—all perfect. We even figured out a way to mold them from Zytel® nylon resin for high strength, limited-cycle uses. The mating surface is a continuous loop. Press the two together and they hold with astonishing force.



Ideas are everywhere, waiting to be found and transformed.

It's in the bag

Some of the best ideas are stimulated by simply restating an existing situation as a problem.

Here's a case in point. Indulge us by thinking about meat carcasses. They're shipped in refrigerated cars and they can last two or three weeks.

The industry was well aware of the spoilage situation but had always considered it a fact of life. Until an ingenious packaging company challenged those long available facts and decided that there was indeed a problem: how to extend the shelf life of fresh beef from two weeks to two months.

The solution was to exclude oxygen from contact with the meat. So the company decided to cut the meat into 75-pound pieces, put them into plastic bags, pull a vacuum and seal the bags. Presto. Now the meat remains fresh for one to two months. And it's in better condition... better in looks, smell, taste, tenderness. Delicious!

The critical factor is the bag. Nylon (for toughness, abrasion-resistance and oxygen barrier) is co-extruded with film of Surlyn® ionomer resin (for a good heat seal) to produce a meat packaging film with a host of packer, shipper and consumer benefits.

To give *you* food for thought, "Zytel" nylon can be extruded on many substrates ranging from paper and aluminum to a potpourri of plastics. Perhaps the case of the problem-that-wasn't will inspire you to look askance at "accepted facts" and redefine them as problems looking for a solution.

Invitation to a Dialog

We hope this treatise has encouraged you to stretch your mind in different directions in your own field. Perhaps you can borrow an idea from nature or restate an existing situation as a problem. Well, let it happen.

Here is a simple way for you to gain the benefits available from our technology. If our expertise in the properties and uses of engineering plastics may be of help, you can continue this dialog by either writing Dick Johannes, Du Pont Company, Plastics Department, Room D-13064, Wilmington, DE 19898 on your company letterhead, or calling him at (302) 774-5826.

DU PONT

Established 1845 SCIENTIFIC AMERICAN August 1975 Volume 233 Number 2

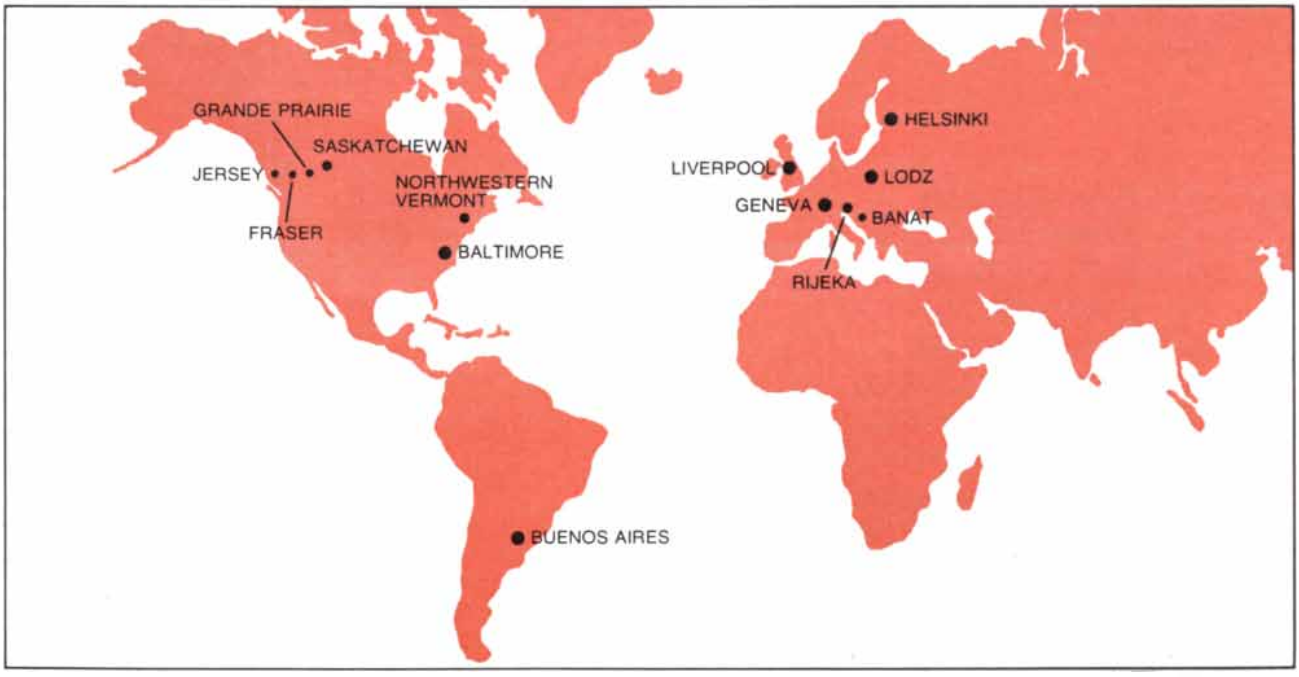
International Comparisons of Medical Care

In the planning of health-care systems it would be useful to know if there are universal patterns in their use. Such patterns are sought in a survey of health care in 12 areas of seven countries

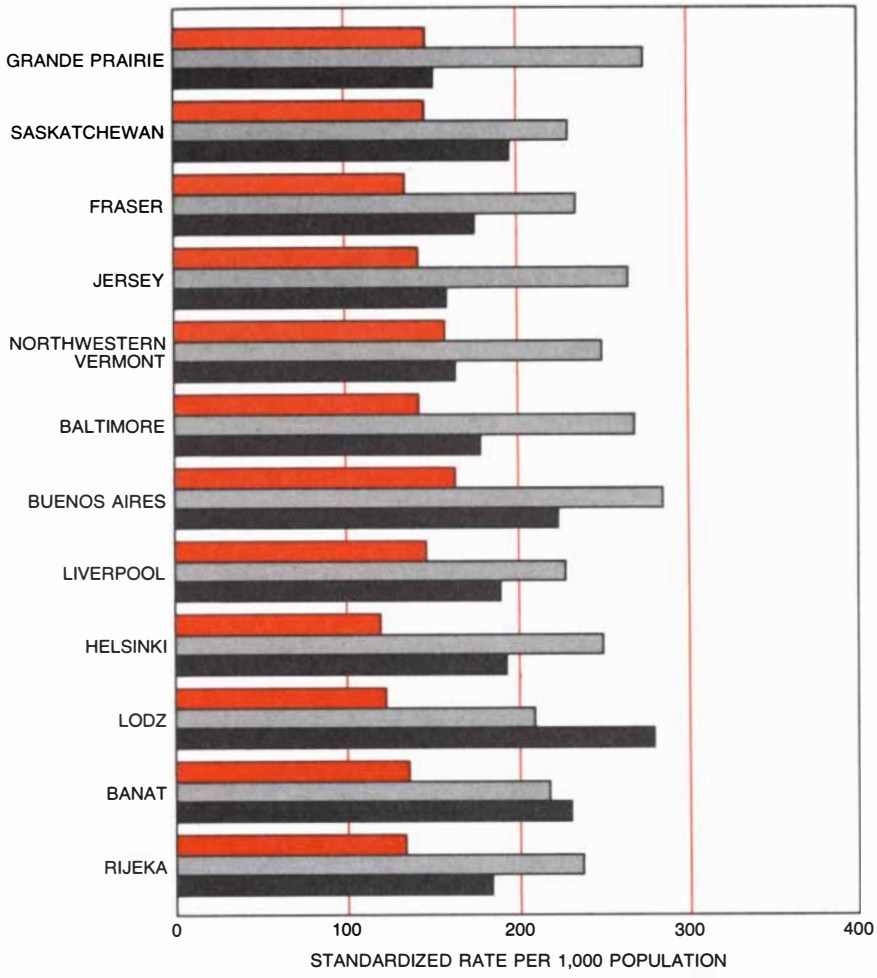
by Kerr L. White

Few social services are more difficult to plan than the provision of health care for entire populations. From country to country and also within a country one finds large variations in the demand for health services and in the allocation of resources for meeting the demand. The issue of achieving a rational equilibrium between demand and resources becomes more acute as the cost of all the constituents of health care goes up, as the variety of medical services that can be provided increases and as the amount of public money spent to support the system rises.

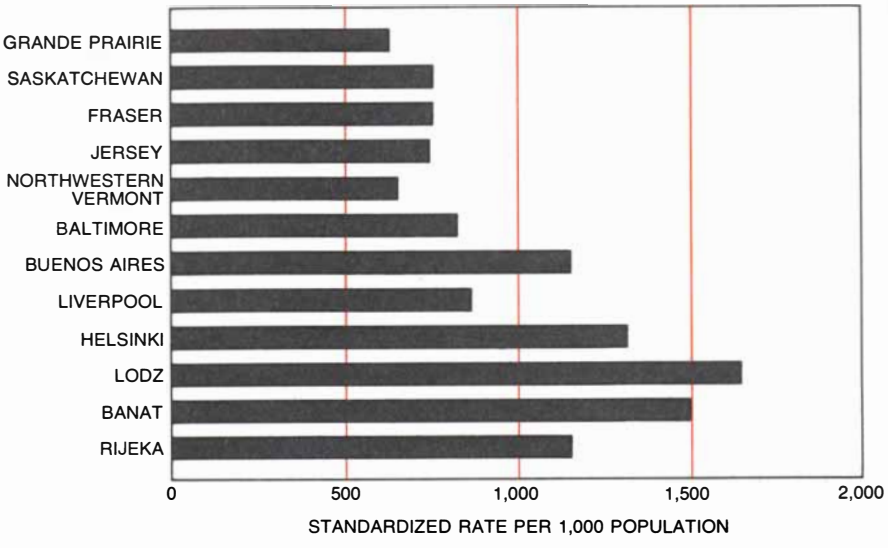
What is needed in confronting this problem is a strong base of information and a better understanding of fundamental relations among the components of a health-care system. This article describes an effort to build such a base by means of a comparative study on a large and international scale. The study shows that epidemiological and statistical sur-



AREAS OF STUDY are indicated on this map. Four of the 12 areas surveyed were rural (*small dots*), three semirural (*medium dots*) and five urban (*large dots*). Study was done under the auspices of the World Health Organization, which has its seat in Geneva.



STATE OF HEALTH in the 12 study areas is depicted in terms of "healthy" and "functionally healthy" people and those who reported sick days. As defined by the study, healthy people (*color*) reported none of the indicators of ill health employed in the study. Functionally healthy people (*gray*) reported one of the conditions or more but were not in bed or restricted in activity. People with sick days (*black*) were in bed or restricted in activity during a recall period of two weeks. The medians were 143, 244 and 184 per 1,000 population.



EXTENT OF ILLNESS is compared in terms of the volume of sick days or social dysfunction in each population during a two-week period. The median was 845 per 1,000 population.

vey methods can provide useful information about health-care systems that is not available from other sources. These methods have been shown to be applicable in industrialized societies; in metropolitan, semirural and rural areas; in localities with a small population and those with a large population, and in places where health-information systems are sophisticated and places where such systems are comparatively modest. A similar approach, perhaps with modifications, can be applied in developing countries where large-scale information systems based on records are unlikely to materialize for decades.

The study is designated the World Health Organization / International Collaborative Study of Medical Care Utilization. It seeks to enhance understanding by pooling and comparing the experience of different population groups in different parts of the world. It has involved a decade of effort by some 90 professional people and some 300 technical workers in 12 study areas in seven countries: four study areas in Canada, two in the U.S., two in Yugoslavia and one each in Argentina, Finland, Poland and the United Kingdom. It has developed information about these health-care systems that does not necessarily support conventional conceptions, as American readers in particular will realize when they compare the findings for the two U.S. study areas with those for other countries.

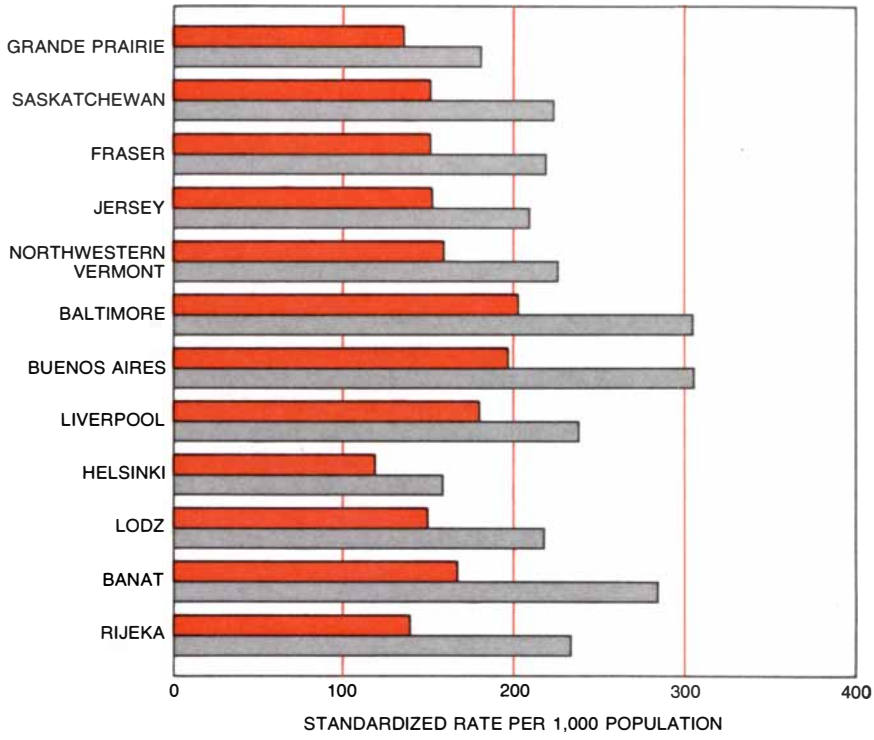
Although the study was international, the reader should be mindful that the 12 study areas are not necessarily representative of their countries or political subdivisions and were not meant to be. We sought instead to choose areas where there were groups of interested investigators with the required research capability and with access to geographically defined populations that could be randomly sampled and whose health-care resources could be determined.

In Canada the study areas were Grande Prairie, which is a rural county in Alberta; Saskatchewan, meaning in this connection the city of Saskatoon and certain urban and rural municipalities centered in North Battleford in the province of Saskatchewan, and two semirural communities, designated Fraser and Jersey, in British Columbia, which are similar except that Fraser is appreciably better off than Jersey in terms of health manpower. In the U.S. the study areas were a five-county region in northwestern Vermont, and the city of Baltimore and its five surrounding counties, both of which had previously been desig-

nated by the Federal Government for statistical purposes. In Argentina the study area was the federal district of Buenos Aires, including the city proper and 18 surrounding municipalities. In the United Kingdom the study area was designated Liverpool but actually consisted of a considerable part of the counties of Lancashire and Cheshire in addition to the city of Liverpool. In Finland the study was carried out in the Helsinki metropolitan area. In Poland the study area covered four separate communes in the province of Lodz, including two sections of the city of Lodz. In Yugoslavia the study areas were Banat, a historically well-defined rural region northeast of Belgrade, and Rijeka, where the territory covered is both urban and rural. Broadly speaking, the study embraced five large metropolitan areas (Baltimore, Buenos Aires, Liverpool, Helsinki and Lodz), four rural areas (Grande Prairie, Fraser, Jersey and Banat) and three areas that are partly urban and partly rural (Saskatchewan, northwestern Vermont and Rijeka).

Our primary source of data was a household survey of the noninstitutionalized population (that is, people not in hospitals or long-term care facilities at the time of interview) by means of interview questionnaires, which were administered in 1968–1969 in four successive quarters in interviews with probability samples of households in each study area. The rate of response was unusually high, ranging from 90 to 99 percent. The number of people in the sample was approximately 48,000; they represent more than 15 million people living in the 12 study areas. Because comparability of methods is essential for a study of this kind, elaborate measures were taken to translate and retranslate the questionnaires into the five major study languages (English, Spanish, Finnish, Polish and Serbo-Croatian), to prepare nine standard procedural manuals for the fieldwork and analysis and to train all fieldworkers in the uniform application of the methods used. The full report of the study is being published in a book, *Health Care: An International Study*.

The results summarized here are confined to major findings about medical care, although the study also covered dental care and visits to eye specialists and to health workers other than physicians (podiatrists, nurses and so on). The study did not address three important considerations of the health planner: the efficacy of the different types of diagnostic and therapeutic intervention employed, the effectiveness with which they were provided to people who would benefit from them and the efficiency with which they were administered. These important measures of the quality of health care are essential components in the political equation that seeks to balance equity of access to the health-care system and control of the system's cost. The fact that the study omitted them reflects the extreme difficulty, particularly in an international study, of measuring aspects of this kind rather than any failure to recognize their significance.



RELIANCE ON PHYSICIANS by the 12 populations is portrayed in terms of individuals who saw a physician within two weeks (*color*) and in terms of the volume of contacts with physicians during the period (*gray*). Median rates were 155 and 225 per 1,000 population.

Before examining variations in resources and variations in how they are employed, it is important to know whether the study areas exhibit gross differences among the 12 populations in the major indexes of health (or the absence of health) when the effects of the different age and sex structures of those populations are removed by standardization of the rates for age and sex. In general the populations do not differ materially in health status as judged by such traditional indexes as mortality rate (from 6.9 to 9.2 deaths per 1,000 population) and proportional mortality, meaning the percentage of deaths of people aged 50 or more out of total deaths (from 74.2 to 88.7 percent). Only infant mortality rates show a wider range (from 15.8 to 47.1 deaths of children less than a year old per 1,000 live births). On this measure Buenos Aires, Lodz and Banat have substantially higher rates than the other nine areas.

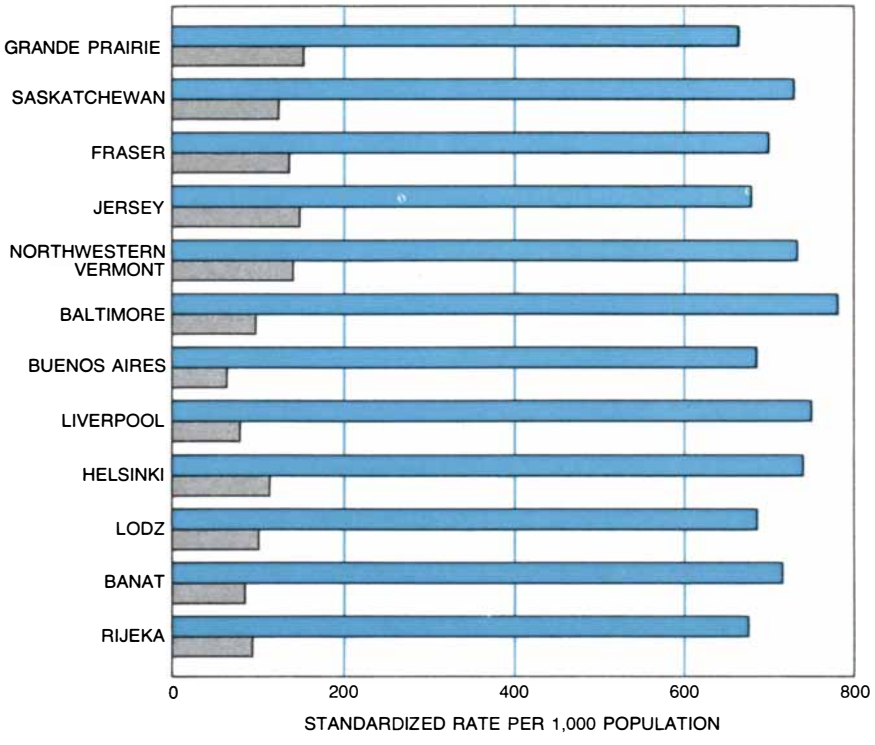
It is also interesting to note the consistency among the 12 populations in the number of people who are considered "healthy" or "functionally healthy" under the definitions employed in the study. Essentially a person was defined as being healthy if he did not report (at the time of interview or for a specific period of recall preceding it) any of five indicators of ill health. These partly overlapping indicators are social dysfunction, meaning days spent in bed or in restricted activity because of illness; perceived morbidity, meaning illness that the respondent acknowledged during an interview; symptoms of specific psychobiological dysfunction (phlegm and cough, chest pain, shortness of breath, pain, stiffness or swelling of joints and anxiety); perceived dental morbidity, and perceived visual morbidity. The functionally healthy included people who, although they did not report days in bed or days of restricted activity, responded affirmatively to one question or more concerning the presence of a slight chronic but not disabling condition, showed the lowest level of one or more of the five symptom sets of psychobiological dysfunction or of dental

or visual morbidity or had visited a physician for a reason not associated with a high level of pain or distress.

In the study areas the group of "healthy" people, consisting mostly of children and young adults, showed a fairly narrow range from 120 to 164 per 1,000 population. The "functionally healthy" ranged from 210 to 286 per 1,000. Taken together, the respective medians of 143 and 244 suggest that about four in every 10 people were not impaired in social functioning at the time of the interview or just before it, according to the indicators employed in the study. One could argue that the stability of these measures across 12 such varied study areas contradicts the widely held view that the demand for medical care is fed from a bottomless pit of need. Evidently the potential demand is limited in each of the 12 populations.

As one might expect, more variability is observed when the perspective is changed to compare measures of disability among the people (roughly six in 10) who did not emerge as healthy or functionally healthy under the terms of the study. One of the measures is sick days, meaning days when the respondent was in bed or was restricted in his activity because of illness. The range for individuals is almost twofold, from 153 per 1,000 in Grande Prairie to 280 per 1,000 in Lodz. The rates for the total number of sick days reported vary even more, from 634 per 1,000 population in Grande Prairie to 1,650 per 1,000 population in Lodz. On this measure of potential demand the rates are significantly higher in Buenos Aires and the four study areas in continental Europe than they are in the six North American areas and Liverpool. The reasons are difficult to determine. In part they may reflect cultural differences in the manifestation of sickness rather than differences in illness as determined clinically.

A different but overlapping measure of the potential pressure on a health-care system is the presence in the population of "chronicity with disability," which is defined as a physical impairment or handicap or a long-standing health problem with some resulting disability. Here the range is fourfold (from 77 to 304 per 1,000 population). The rates are relatively low and consistent in all except the four areas in continental Europe. They are particularly high in Lodz and the two areas in Yugoslavia, perhaps reflecting the exposure of those areas to the direct ravages of two world wars. More detailed statistics show that the major incidence of this measure of ill health is among the adult population, except that in Banat it is also high among the children. The higher rates in the four continental European areas persist through all groupings of adults by age and sex.



USE OF PHYSICIANS AND HOSPITALS over a period of 12 months is charted. The bars show rates at which individuals had at least one contact with a physician (*color*) and one overnight admission to a hospital (*gray*). Medians were 709 and 108 per 1,000 population.

The measures described can be employed in calculating what we term a dependency ratio for each study area. It assumes that all children, all those aged 65 or more and all those reporting sick days or chronic illness with disability require some measure of support and care by the healthy and functionally healthy, presumably independent and probably productive people aged 15 to 64. With such a dependency ratio one can compare across study areas the extent to which productive energy may be offset by the less productive or nonproductive members of the society. The ratio of the dependent population to the others is particularly high in Lodz and the two study areas in Yugoslavia. Taking all the study areas, the ratio is 1.3 : 1, which is to say that for every three healthy or functionally healthy adults four dependent people are being supported or cared for because of youth, old age or illness.

This ratio reflects the relative balance between dependent and independent people. It says nothing, however, about the proportion of the independent group's energy and money spent in support of the dependent group, and it does not pretend to indicate whether such relations are more or less desirable or better or worse than others. Nonetheless, such ratios and their monetary equivalents do have implications not only for the organization of health services but also for the provision of other social services, for programs designed to promote security of income, for tax policy and for the productivity of the society.

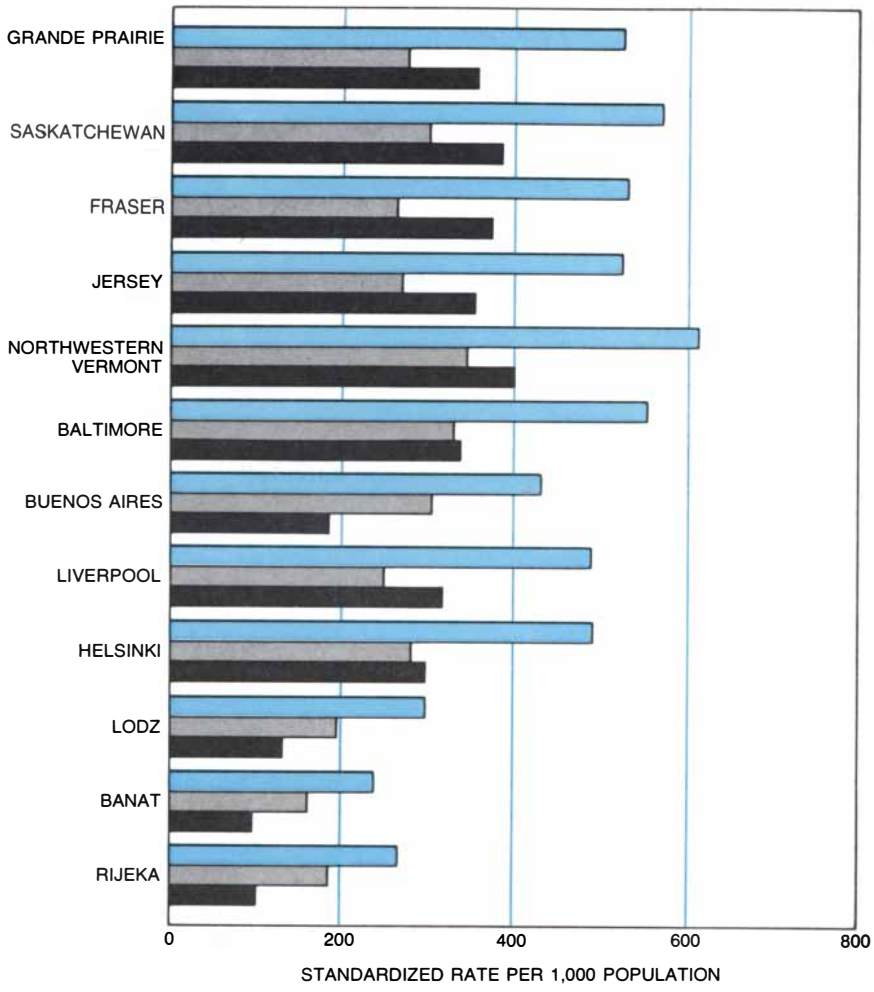
Before turning to the study's more specific findings on how the people in the 12 study areas make use of physicians, hospitals and medicine, the reader should be reminded that any system of health care involves complex interactions of people's perceived need for care and the allocation of resources to meet that need. Many choices can be made in achieving a balance of facilities and manpower. Moreover, how much need is translated into the actual use of services reflects the availability and accessibility of resources. Comparative studies of how different populations employ such resources as physicians, hospitals and medicine provide a useful point of departure for understanding these complex issues. Our study examined both the extent to which individuals called on these re-

sources and the types and amounts of the resources that were deployed. In this way we could see not only the relation of individuals to the health-care system but also the extent to which the society at large was relying on the system.

One of our measures of the relations between individuals and physicians is the number of people in the sample who went to a physician (or called one in) at least once during a two-week period and during a 12-month period. These were "ambulatory care" encounters, which means that the patient was not hospitalized at the time of contact. For the two-week period the rate ranged from 120 per 1,000 population in Helsinki to 197 in Buenos Aires and 202 in Baltimore, with a median of 155 per 1,000. The rates for the five rural and semirural areas in North America and for Lodz, Banat and Rijeka tended to be lower; the rate for Liverpool was intermediate. A similar stability was observed for individuals seeing a physician within 12 months; the range was from 663 to 779 per 1,000 population, with a median of 709. In short, about seven people in every 10 see a physician at least once a year in most of the study areas.

The figures on the two-week rates for the volume of contacts with physicians are also instructive because so little is known about how often a patient should or might consult a physician once the initial contact has been made. These rates for the total number of contacts with physicians per 1,000 population show somewhat more variation than the rates for individuals with contacts within two weeks. The range is from 158 contacts per 1,000 population in Helsinki to 306 in Buenos Aires; the median is 225 per 1,000.

None of the three measures provides any obvious evidence for the widely held view that the mechanisms for paying the physician and the presence or absence of financial barriers significantly influence the rate at which people call on the services of physicians. For example, one could argue that the fee-for-service method of payment in northwestern Vermont, Baltimore and Buenos Aires may lead physicians to encourage more discretionary visits in those areas; on the other hand, it could be argued that the absence of financial barriers (because of a national health service or national health insurance) may account for a somewhat similar effect in Liverpool, Lodz and the two study areas in Yugoslavia by leading patients to see physicians more frequently.

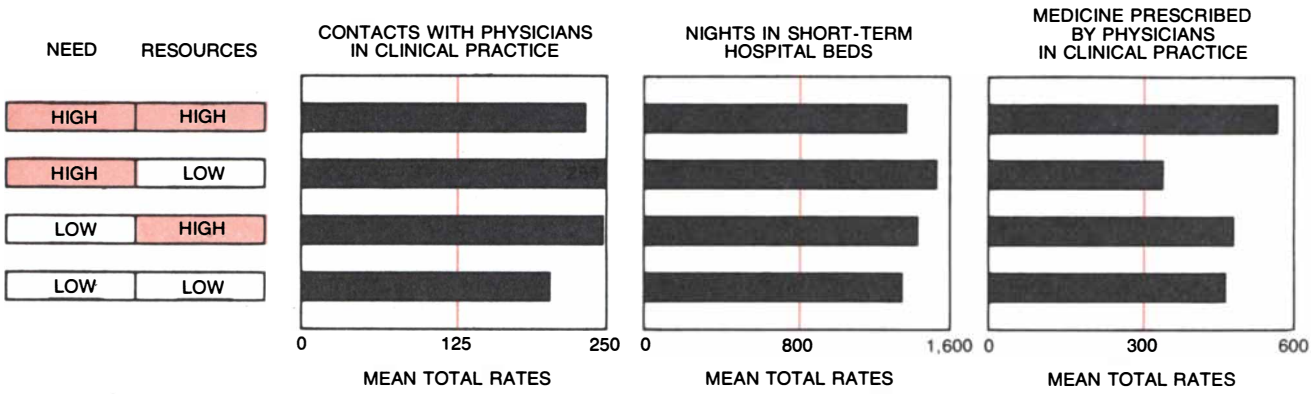


RELIANCE ON MEDICINE in the 12 populations is shown in terms of the rate at which people took any medicine within two weeks (*color*) and, if so, whether it was prescribed medicine (*gray*) or nonprescribed medicine (*black*). Some people may of course have been taking both kinds of medicine. The medians were 507, 271 and 327 per 1,000 population.

Variations in the supply of physicians are likewise not sufficient to account for the differences observed; Buenos Aires and Helsinki have the two highest ratios of physicians in clinical practice (19.7 and 12.2 per 10,000 population respectively), but the coverage by health insurance is broader in Helsinki. There it appears that neither low financial barriers nor a supply of clinical physicians above the median is associated with the rate at which individuals call on physicians in a two-week period or with the volume of calls on physicians in two weeks. On the other hand, the rate at which individuals see physicians over the 12-month period is higher in Helsinki than it is in Buenos Aires. The contrast suggests that the organization of the system and the mechanisms of payment may moderate discretionary visits, at least in the long run. These variations illustrate why it is necessary to look at the total picture of resources—physicians and nonphysicians, generalists and specialists, ambulatory care as against hospital care—and at financial and behavioral patterns in predicting the use of the services of physicians.

A more informative way of examining the availability of services by physicians and their use is controlling for the different levels of perceived morbidity, that is, holding them constant so that utilization can be compared from a consistent base, much in the same manner that age and sex are standardized for the comparison of different population groups. One way we applied such a control was to standardize the level of severity of illness across the study areas; another was to employ a uniform level of chronicity. The patterns that emerge reveal substantial variations in the rates at which populations call on the services of physicians.

When the severity of perceived mor-



DEMAND FOR HEALTH CARE is related to various patterns of need and resources found in the 12 study areas. The bars at left describe patterns of need and resources that were either high or low, meaning above or below the median for the 12 areas. Reading to the right in each case, one sees how the pattern of need and resources compares with use of physicians, hospitals and medicine.

bidity is standardized, it appears that children are given a higher priority than adults in the two Yugoslavian areas and, to a lesser extent, in Baltimore, Buenos Aires, Liverpool and Lodz. In other words, in those places larger proportions of children than of adults see physicians. Taking the total population, the highest rates are found in the three largest metropolitan areas (Baltimore, Buenos Aires and Liverpool), which have quite different systems of health care.

On the other hand, the standardization of chronicity tends to stabilize the rates. (Chronicity is a measure of the incidence of long-standing health problems such as a physical impairment or a chronic illness.) Given similar burdens of chronicity, the populations of the 12 study areas are seen to call on physicians at least once a year at similar rates.

Another way of assessing the impact of perceived morbidity on the use of physicians is to compare the number of contacts within two weeks by people who have perceived themselves as being at the highest level of severity of illness as measured by a "bother, hurt or worry" index. The measure should be fairly discriminating, since it reflects the resort to physicians only by people who reported a high level of perceived acute illness at the time of the visit. On this basis the patterns within countries (Canada, the U.S. and Yugoslavia) are similar, but large differences appear between the six study areas in North America and Buenos Aires and the five European areas. Liverpool has the highest rate of contacts with physicians, and the four areas in continental Europe have by far the lowest rates. Evidently financial barriers play only a small role at times of acute illness, since the individual's liability for paying the cost of health care is minimal in all the European study areas and from moderate to considerable in Buenos Aires, the U.S. and (at least at the time of the survey in 1968–1969) the four Canadian study areas.

One can also distinguish between discretionary care initiated by the patient (and largely influenced by the physician thereafter) and the demand for care that is generated by administrative requirements in the form of physical examinations for such purposes as school, employment and insurance. The rates for people whose most recent physical examination within the 12 months preceding the interview was attributable to an administrative requirement illustrate that this form of care can place a substantial burden on health resources. The rates are lowest in the four Canadian areas, close to the median in Buenos Aires and the two study areas in the U.S., considerably higher in Liverpool and Helsinki and highest by far in Lodz and the two study areas in Yugoslavia. The range is from 203 to 759 such examinations per 1,000 population; the median is 328. Here is a large demand for a procedure that is of doubtful efficacy in many instances and that represents a questionable use of medical resources in the face of their relative scarcity.

Although ambulatory care by physicians may offer the best prospect for the prevention and early detection of disease, it is the hospital care of inpatients that consumes both the bulk of the society's capital investment in health-care facilities and the major part of the cost of medical care. Our study's examination of this component of the system reveals the interplay between parts of the system and also the way in which the different study areas allocate their resources. The rates for individuals having an overnight admission to a hospital within 12 months show more variation than the rates for individuals visiting physicians, and the pattern is quite different.

For example, Baltimore, Buenos Aires and Liverpool were shown to have the highest rates for individuals seeing physicians within two weeks; in contrast, those areas are among the lowest in the number of people having an admission to a hospital for overnight or longer within a period of 12 months. The range is from 65 per 1,000 population in Buenos Aires to 156 in Grande Prairie, with a median of 108. The rates in Baltimore and the five European study areas are low compared with the rates in the five semirural study areas in North America. Helsinki, which had the lowest rate for individual calls on physicians within two weeks, is highest among the European areas in hospital admissions.

Measuring the volume of nights in hospitals, one finds a threefold range (from 732 to 2,399 per 1,000 population per year) and a median of 1,480. The differences among the study areas may well be associated partly with the availability of hospital beds and with the balance between short-term and long-term beds. Buenos Aires has the lowest supply of hospital beds (4.6 per 1,000 population) and the lowest rates for both measures of the use of hospitals. Grande Prairie has a high ratio of beds to population (12 per 1,000), particularly short-term beds (7.3 per 1,000), and the highest rate for individuals with overnight admissions; except for Lodz it also has the highest rate for volume of hospital nights. Although Lodz is below the median in both total beds and short-term beds, it has by far the longest mean length of stay and therefore the highest rate for hospital nights. This picture

might be attributable to any of several things: patterns of practice, a higher proportion of patients having serious illness and therefore requiring longer stays in hospitals, or relatively low efficiency of hospitals.

In view of these patterns we believed the amount of unmet need for the services of physicians might have a material influence on the total of hospital nights. We measured the unmet need in terms of the percentage of people who, although they reported illness of high severity within two weeks and the desire to consult a physician, did not actually obtain the services of one. Their inability to see a physician could have been attributable to any of several reasons, such as a relative lack of clinical physicians or a relative imbalance in the types of practitioners. Our measure did reveal a correlation: Where the apparent unmet need for physicians is highest the total of hospital nights is also highest. The relation is observed even with the exclusion of Buenos Aires, which has by far the lowest ratio of short-term beds to population among the study areas and also the lowest rates of use of hospitals.

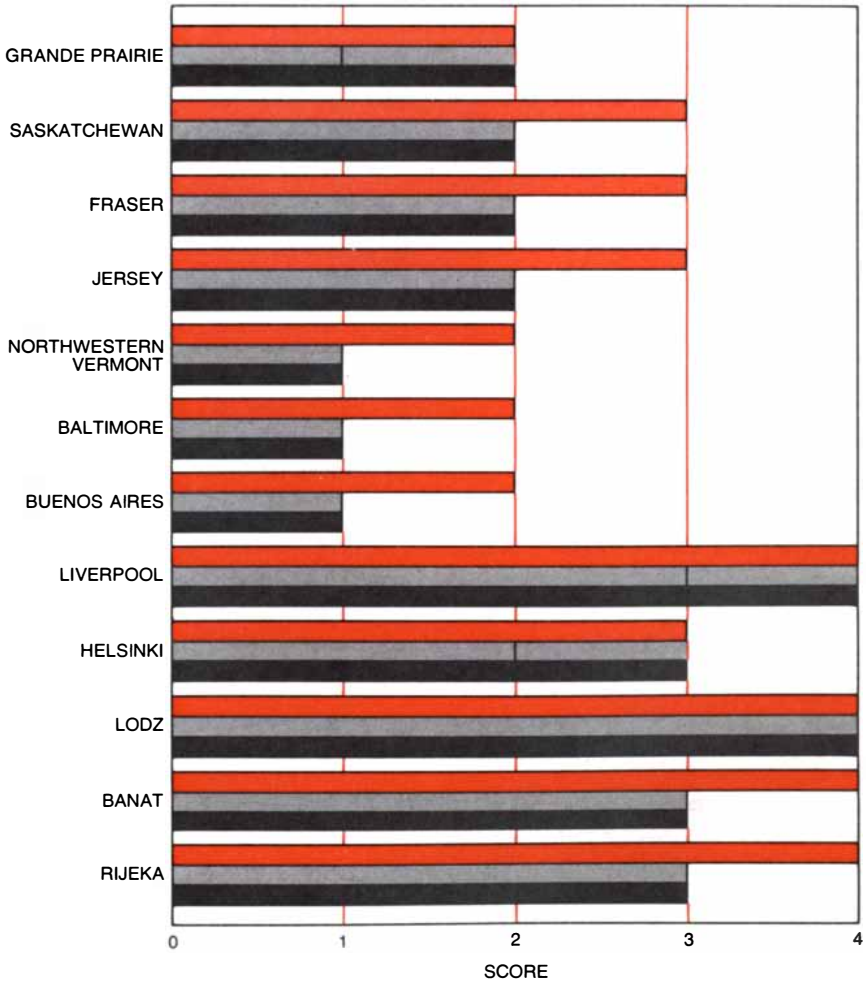
It is of interest here that the countries with two or more study areas show similar patterns. The suggestion is that in relation to the demand for medical care the balance between physicians (perhaps also types of physicians) and hospital beds may be a more important determinant of the use of hospitals than the mere availability of beds. The study areas with a comparatively high ratio of beds do not seem to be the most responsive to this proxy measure of the pressure on health services. Perhaps, then, what is regarded as an excess capacity of beds is not the major determinant of the use of hospitals, and a more important factor may be the availability and accessibility of physicians in ambulatory care.

In examining the use of medicine we asked each respondent whether he had taken any medicine during the two days before the interview and, if he had, what the division was between prescribed and nonprescribed medicine. The findings showed that the study areas where the overall use of medicine is relatively high are likely to exhibit equally high use of both prescribed and nonprescribed medicine. Patterns within the countries tend to be similar, but cross-national differences are substantial and consistent. For example, considering the two areas in Yugoslavia and the two in the U.S. as two groups, one finds a twofold difference from low (Yugoslavia) to high (U.S.) in the consumption of medicine. Liverpool and Helsinki are at intermediate levels, and the four Canadian areas are closer to the two in the U.S.

The differences are most marked for people taking nonprescribed medicine: the range of rates is fourfold. It is hard to believe that the availability of medicine of established efficacy differs widely among these seven industrialized nations. Although cultural attitudes may influence the faith people have in nonprescribed medicine and even the extent to which people report taking such medicine, it seems probable that the dominant influences are the prescribing habits of physicians, the distributional and promotional efforts of pharmaceutical manufacturers and the financial arrangements of the health-care system. Affluence also seems to push up the consumption of nonprescribed medicine.

Further evidence supporting the impact of cultural and promotional influences on the consumption of medicine is provided by the rates for people who were taking medicine even though they reported neither illness nor a chronic health problem. The data indicate that much of this consumption consists of vitamins, tonics and laxatives. The resort to medicine by what could be as much as a third of this group of presumably healthy people raises questions about the perceptions they and perhaps their physicians have of health and well-being.

Comparative data of the kind assembled by our study can be combined in a number of ways to elucidate patterns and relations in health care. For example, one can examine the balance of need, resources and use in the light of the social and political characteristics of the study area that influence the various administrative and planning mechanisms



GENERAL ATTRIBUTES of the health-care systems in the 12 study areas are scored according to three sociopolitical characteristics, "Health as a societal value" (*color*), "Collectivism versus individualism" (*gray*) and "Distributional responsibility" (*black*), as defined in the text. In the scoring system 1 is low, 2 is intermediate, 3 is high and 4 is maximal.

whereby priorities are controlled and resources are allocated. The step can be taken by assigning approximate weights to attributes of a health-care system that reflect the extent of formal control and planning and then calculating scores for them.

The attributes we chose are "Health as a societal value," "Collectivism versus individualism" and "Distributional responsibility." The first attribute reflects the extent to which health is a subject of concern on the part of the society and receives priority in the allocation of the society's total resources. The second establishes a scale for the emphasis the society gives to the control of the distribution of its health-care measures and resources. Distributional responsibility reflects the extent to which decision making is centralized or decentralized.

The relations are depicted in the illustration on the preceding page. As one might expect, the five study areas in Europe can be seen to have the highest scores on the three attributes, with Buenos Aires and the two areas in the U.S. showing the lowest scores. A companion illustration on these two pages portrays, in terms of percentages of the medians, the extent to which this speculative but indicative analysis corresponds to the perceived need, the available resources and the reported use.

The analysis gives rise to intriguing questions. For example, in terms of the measures employed in this study the Saskatchewan, Fraser and Jersey areas could be regarded as being reasonably balanced; the characteristics of their health-care systems are at intermediate levels. On the other hand, Grande Prairie, which is in many ways the most rural of the four study areas in Canada, has lower relative levels of need, a greater imbalance in the relative supply of physicians and short-term beds and a lower score on the general system attributes described above than the other three Canadian areas have. Are the arrangements in Jersey and Fraser, then, the result of market forces and an open system, or have they evolved in accordance with some informal plan? Is the balance observed for the Saskatchewan area the result of more specific planning or less?

Liverpool is another study area that displays a considerable balance of need, resources and use. Is its relatively high level of volume of contacts with physicians, which is in contrast to a relatively low supply of physicians, a manifestation of greater productivity, higher efficiency, different forms of care or the absence of financial barriers? The question takes on added pertinence when Liverpool is compared with Buenos Aires and the two study areas in the U.S., where the supply of physicians is considerably higher. In Liverpool the ratio of general practitioners to specialists is 1.2 : 1, whereas it is .3 : 1 or less in the other three areas. Is the mix of generalists and specialists as important as the total supply of physicians, or perhaps even more important?

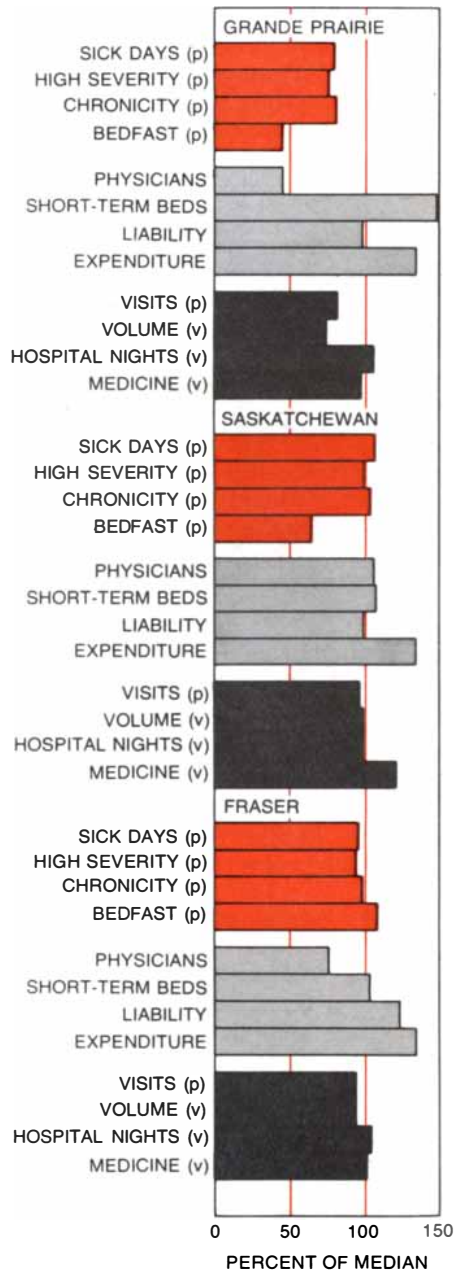
In Helsinki the level of use of hospitals is high in relation to the level of use of physicians, whereas in Liverpool the situation is reversed. Do these relations result from policy on the allocation of resources? Which arrangements are the most effective economically and clinically?

Among the 12 study areas the relative level of resources in Helsinki is in every instance above the median, and yet the relative levels of use of services are low. These findings are observed in the presence of a relative level of individual liability for meeting the costs of health care that, at the time of the survey, was the highest among the European study areas. Helsinki's score on the general attributes of its health-care system was the lowest among the five.

Banat and Rijeka are trying to bring their resources into balance with their above-average needs. Both of them have high scores on the general characteristics of their health-care systems. Where should the policymakers make their next investments in order to reduce the relatively high levels of perceived need, particularly for the chronically ill? Should the supply of physicians be increased? Should new physicians be encouraged to go into general practice or primary care rather than into specialties? Should more hospital beds be provided? If so, should they be short-term or long-term beds?

Lodz has levels of perceived need similar to those in the two Yugoslavian areas and the highest score of any area for overall characteristics of its health-care system. Do these findings reflect a requirement for the prudent use of resources? (It should be noted that Lodz has the lowest level of health expenditure as a percentage of national income.)

A different kind of exercise shows that these are not abstract questions. It involves comparing the study areas that are at the extremes in resources and use and applying the comparisons to a hypothetical city with a population of one million or to the 200 million people in the U.S. For example, the difference between the minimum supply of physicians and the maximum (Grande Prairie with 6.5 per 10,000 population and Buenos Aires with 19.7) would mean a difference of 1,320 physicians in a population of one million. For a population of 200 million the difference could absorb the annual output of 100 or more medical schools. The figures for hospital beds (Buenos Aires has 46 per 10,000 population and Jersey has 144) are even more dramatic: for a city of one million a difference of 9,800 total beds means a doz-

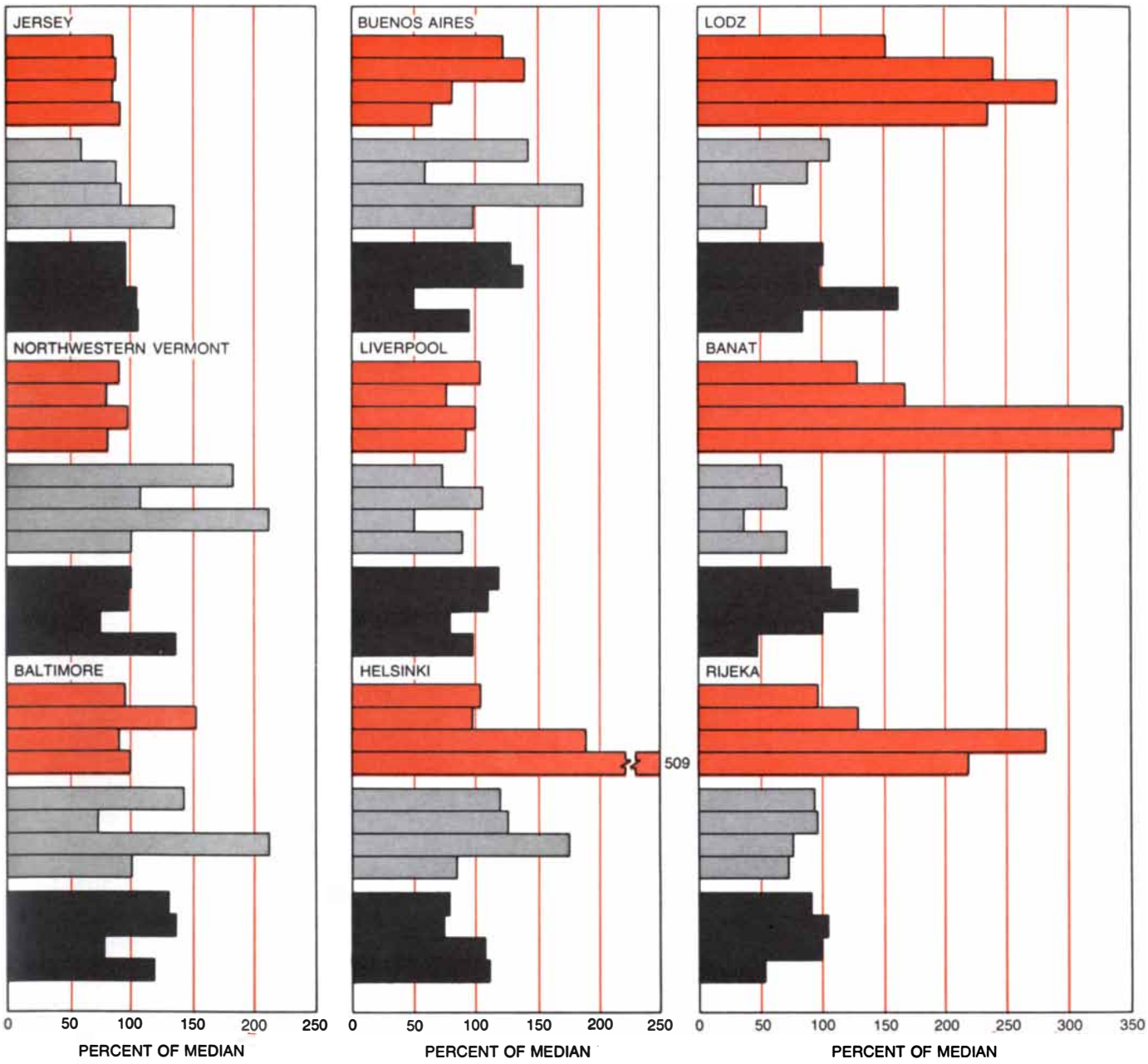


GENERAL COMPARISON is made in terms of each study area's perceived need (*color*), health-care resources (*gray*) and use

en or more large hospitals with 700 or 800 beds each and perhaps as many as 2,000 hospitals for a population of 200 million.

With respect to productivity, a difference of 91,000 people with overnight admission to a hospital during a year is postulated for a city of one million. If they were cared for in the study area with the highest rate of turnover of patients per hospital bed (Jersey, with 46.3 patients per bed per year), it would require almost 2,000 short-term beds, or perhaps three hospitals with about 700 beds each. If the patients were cared for in the study area with the lowest rate (Liverpool, with 20.9 patients per bed per year), about 4,300 beds, or six hospitals of about 700 beds each, would be required. For a country with a population of 200 million the difference would be some 600 large hospitals.

The financial and social investments represented by these figures are huge. Granted that the estimates are conjectures at best, they nonetheless provide an indication of the impact of different mixes of manpower and facilities and of different levels of productivity and efficiency on the use of resources that are scarce and expensive and are likely to remain so in virtually all health-care settings for the foreseeable future. Even if the magnitude of the differences invites skepticism, the fact that differences have been shown to exist should encourage further comparisons in order to obtain more precise estimates of need, resources and use in the context of the organization and the objectives of health-care systems and in relation to competing demands for other social services.



of the health-care system (*black*). Letters *p* and *v* show whether measures are of personal activity or of volume of activity. Under resources "Liability" means individual liability for meeting the cost of health care and "Expenditure" reflects the area's outlay on health care. Under use "Visits" means visits to a physician by individuals and "Volume" means total demand made on physicians' services.

GIANT RADIO GALAXIES

A powerful radio telescope in the Netherlands has revealed that energetic radio sources associated with galaxies are the largest objects known in the universe, extending millions of light-years

by Richard G. Strom, George K. Miley and Jan Oort

One of the most powerful of the present generation of radio telescopes is located at Westerbork in the northeastern Netherlands. The instrument operates by electronically combining the radio signals received by 12 antennas of relatively modest size, thereby simulating the response of a telescope of huge dimensions. Since the Westerbork telescope went into operation in 1970 its users have both discovered and examined a number of remarkable radio galaxies. Some are enormous double radio sources, hundreds of times larger than our own galaxy; others have a tadpole-like structure that is probably formed by a galaxy's hurtling through space and expelling clouds of gas that trail behind it. Such seemingly bizarre objects provide astronomers with exciting opportunities to study the evolution of radio sources and to probe the physical conditions within clusters of galaxies.

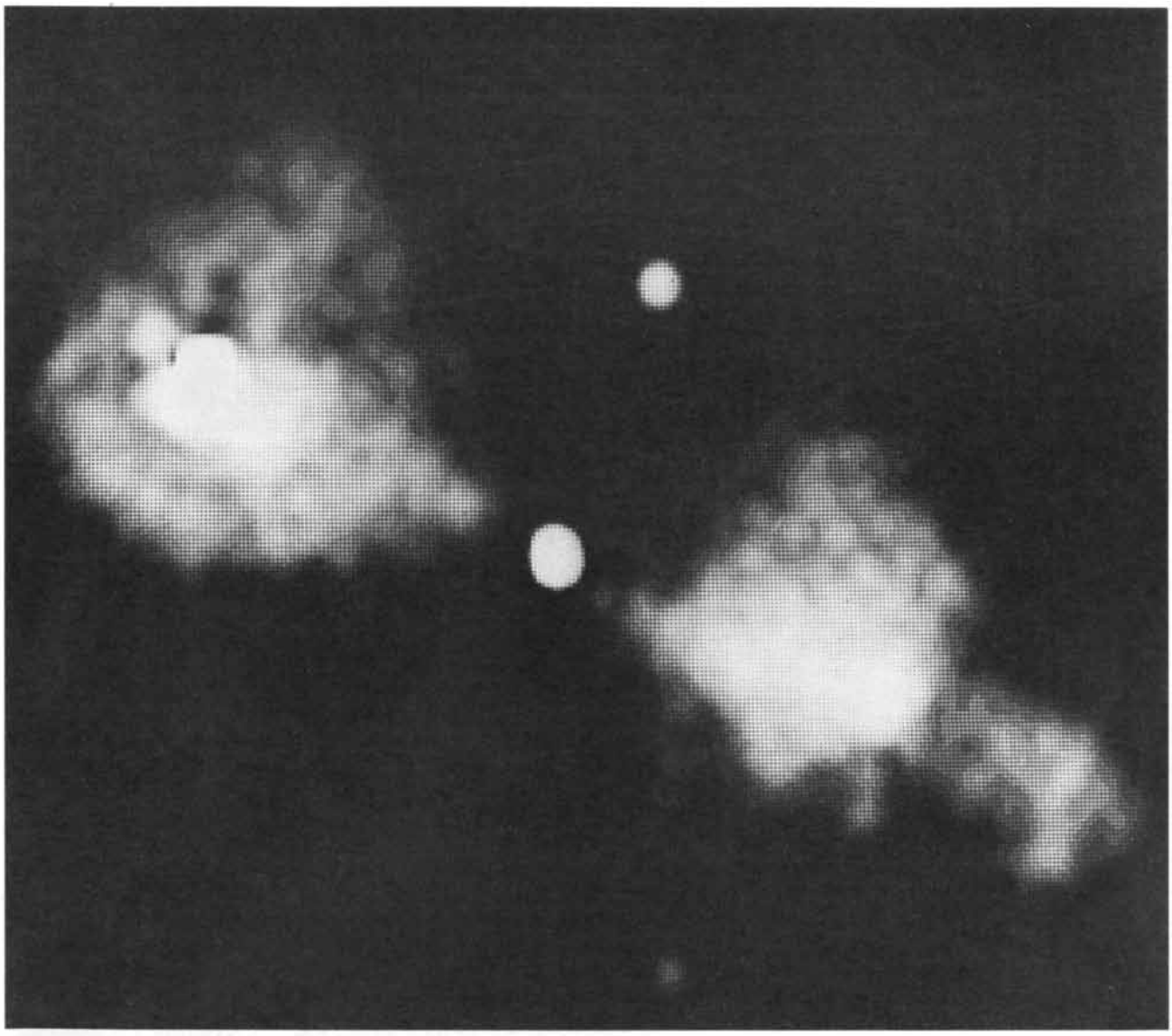
Radio astronomy had a troubled adolescence, when even the type of object being observed was often in doubt. How have the developments of the past 30 years led to instruments such as the one at Westerbork? A central concern has been the angular resolution of radio telescopes, which is proportional to the wavelength of the radiation divided by the reflector's diameter. Radio observations in the 1940's were made at wavelengths greater than one meter with telescopes less than 10 meters in diameter; their angular resolution was thus on the order of 10 degrees, or 20 times the diameter of the full moon.

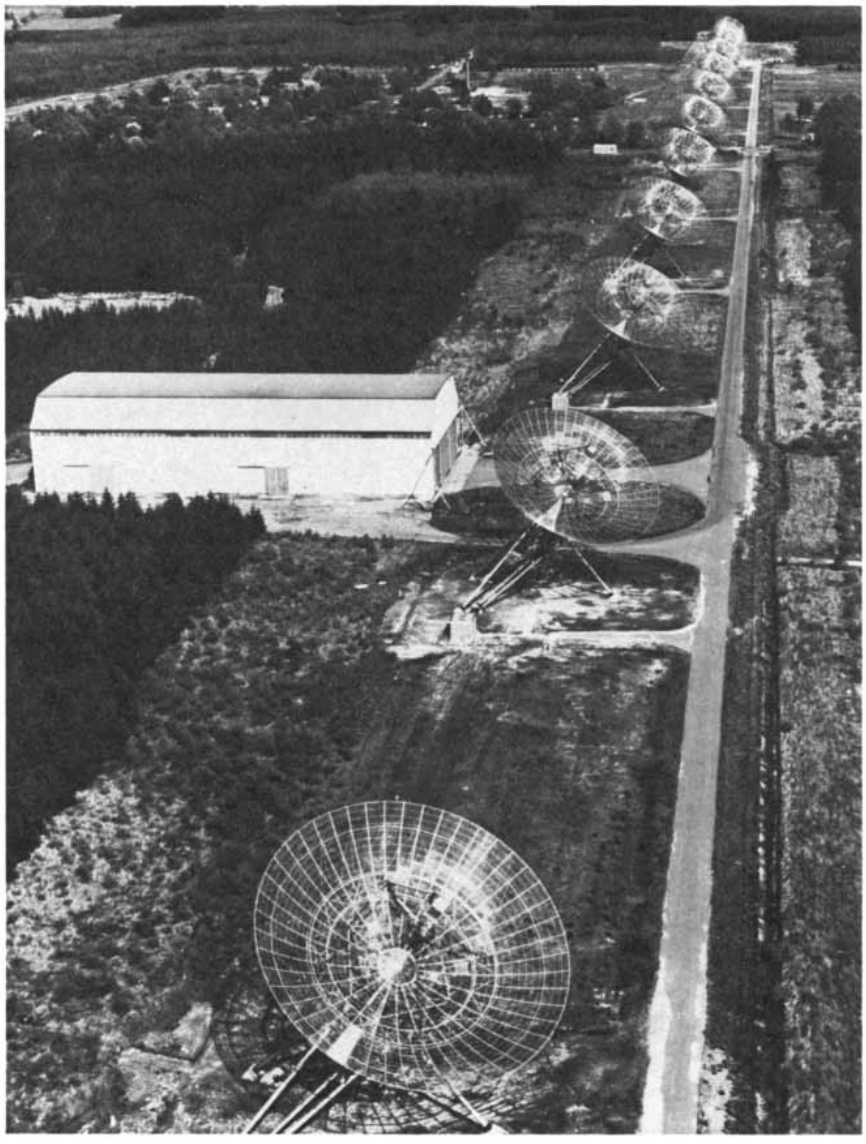
It was possible to join two small reflectors with electrical cables, thereby achieving a resolution equivalent not to the diameter of the individual elements but rather to their separation from each other. It was extremely difficult, however, to keep the signals from the two elements in phase with each other, and it is relative phase that carries all information about the radio source's position. The late Joseph L. Pawsey and his colleagues in Australia ingeniously sidestepped this problem and used the sea as a huge plane reflector, so that the reflected signal from a radio source and the direct signal from the same source would combine to generate interference fringes. Observing from the cliffs above Sydney harbor, John Bolton, Bruce Slee and Gordon Stanley were able to use such fringes to demonstrate that a few strong radio sources are associated with prominent visible galaxies.

After World War II radio astronomy benefited in a number of ways from radar technology. Precision radar dishes abandoned after the war were converted to radio telescopes. With two of these dishes at his disposal, F. G. Smith of the University of Cambridge set about the difficult task of building an interferometer capable of measuring radio positions with an accuracy of one minute of arc. To achieve this he had to set his reflectors 1,000 feet apart, join them with cables, measure their relative position to within an inch and preserve the phase information of the incoming signals. He succeeded in his task and obtained an accurate position for Cygnus A, the strongest radio source in the constellation Cygnus. Training the 200-inch Hale telescope on the same spot, the optical astronomers Walter Baade and Rudolph Minkowski discovered that the radio position coincided with a distant galaxy.

Cygnus A proved to be a watershed

RADIO PHOTOGRAPHS of two galaxies on the opposite page were made with the Westerbork synthesis radio telescope in the Netherlands. The Westerbork telescope consists of 12 dish-shaped antennas, each 25 meters in diameter, arranged in a line 1.6 kilometers long. Its sensitivity and resolution are equivalent to those that would be obtained by a single reflector 1.6 kilometers in diameter (*see illustration on page 29*). In the technique termed aperture synthesis the dishes are paired with one another to form a series of radio interferometers with different base lines. Observations from each of the interferometers are stored on magnetic tape. When all the observations are completed, a computer processes them and constructs a radio map of the object. The map can be displayed on a cathode ray tube so that the brightness of the visual image is related to the intensity of the radio emission. When the image is photographed, the resulting picture shows how the object would appear if the human eye were sensitive to radio wavelengths. The photographic method of representing the data was conceived by Walter Jaffe. The two radio galaxies shown, which are designated DA 240 (*top*) and 3C 129 (*bottom*), are two of the largest objects known in the universe. The double source DA 240 spans 34 minutes of arc in the sky, an angular diameter greater than that of the full moon. At the distance of DA 240 that angular diameter corresponds to a physical diameter of six million light-years. The central radio component of DA 240 coincides with a distant visible galaxy. The intense "hot spot" near the center of the component at left might be the seat of activity feeding energy to a huge cloud emitting energy at radio wavelengths. Double structure is typical of most strong radio sources outside our galaxy. The other object, 3C 129, is a head-tail radio source, so named because of its tadpole-like appearance. The narrow, delicately curved tail stretches over an angular distance of 20 minutes of arc, corresponding to a physical distance of about two and a half million light-years. The head coincides with a galaxy in a cluster of galaxies more than 400 million light-years away. Tail is believed to be a trail of gas left behind by the galaxy as it travels through the cluster. Its distorted shape is possibly due to turbulent motion of intergalactic gas. Ripples around the head are an artifact of the way the image was constructed.





REFLECTORS OF WESTERBORK TELESCOPE lie on a line running east and west. Ten of the 12 reflectors are fixed in position at intervals of 144 meters. The remaining two can be moved along a rail line 300 meters long. The two movable reflectors are at the far end. The Westerbork telescope is operated by the Netherlands Foundation for Radio Astronomy.

for radio astronomy in a number of ways. Although its association with a galaxy was by then not remarkable, two facts set it apart: the galaxy was very distant and the radio power was immense. The generation and release of such huge amounts of energy remain among the most perplexing problems in astronomy to this day. It was immediately clear, however, that similar radio beacons could guide us far beyond the faintest galaxies visible with optical telescopes, thereby serving as probes of space and time. Overnight radio astronomy was changed from being the brainchild of a few interested electronics engineers into perhaps the most important tool of observational cosmology.

Although the cosmological potential of radio astronomy has yet to be fully realized, the study of these radio galaxies over the past 20 years has resulted in a multitude of quite unexpected discoveries. The first surprise concerned Cygnus A itself. At the University of Manchester's Nuffield Radio Astronomy Laboratories at Jodrell Bank in England a new interferometer technique had been developed that made it possible to replace with radio links the cables that would normally connect the individual radio reflectors. Roger Jennison and M. K. das Gupta used this interferometer to produce a result that confounded all expectations: The radio emission from Cygnus A originates in two fairly compact regions that are separated by nearly two minutes of arc and are symmetrically displaced on each side of the visible galaxy. In short, the radio source is double.

It is now clear that double structure is the preferred morphology for strong radio sources outside our galaxy. Although the basic cause of the double structure has so far eluded us, notwithstanding the fact that hundreds of sources have been studied in some detail, we at least have a reasonable understanding of the radio emission itself. The radiation has two remarkable characteristics: it is linearly polarized, that is, it vibrates principally in one plane, and its intensity increases with wavelength. These two properties are the signature of a now well-known process in which radiation is emitted by high-energy charged particles spiraling in a magnetic field. Because such radiation was first observed in high-energy particle accelerators, it was christened synchrotron emission. The way the synchrotron radiation is polarized indicates the direction and uniformity of the magnetic field where it is produced. By studying how the polarized signal varies with wavelength one can deduce whether or not there is ordinary gas permeating the radio source in addition to the particles generating the synchrotron radiation.

Although one of these radio sources may contain as much matter as perhaps a few million suns, that matter is spread throughout a volume so immense as to constitute a vacuum many billions of times rarer than any that can be produced in a terrestrial laboratory. It is a gas with a temperature that may exceed a million degrees Kelvin, laced by electrons, protons and other subatomic particles moving at nearly the speed of light and generating radiation ranging from long radio waves down to millimeter waves, visible light and in rare instances even X-ray emission. Although synchrotron radiation is far stronger at long wavelengths than at short ones, the radiation as a whole is extremely weak. If the largest radio telescope had observed Cygnus A since the beginning of human civilization, the energy collected could power a single half-watt light bulb for no more than a few tenths of a second. Clearly observations of radio sources are severely limited by the insufficient collecting surfaces of our instruments.

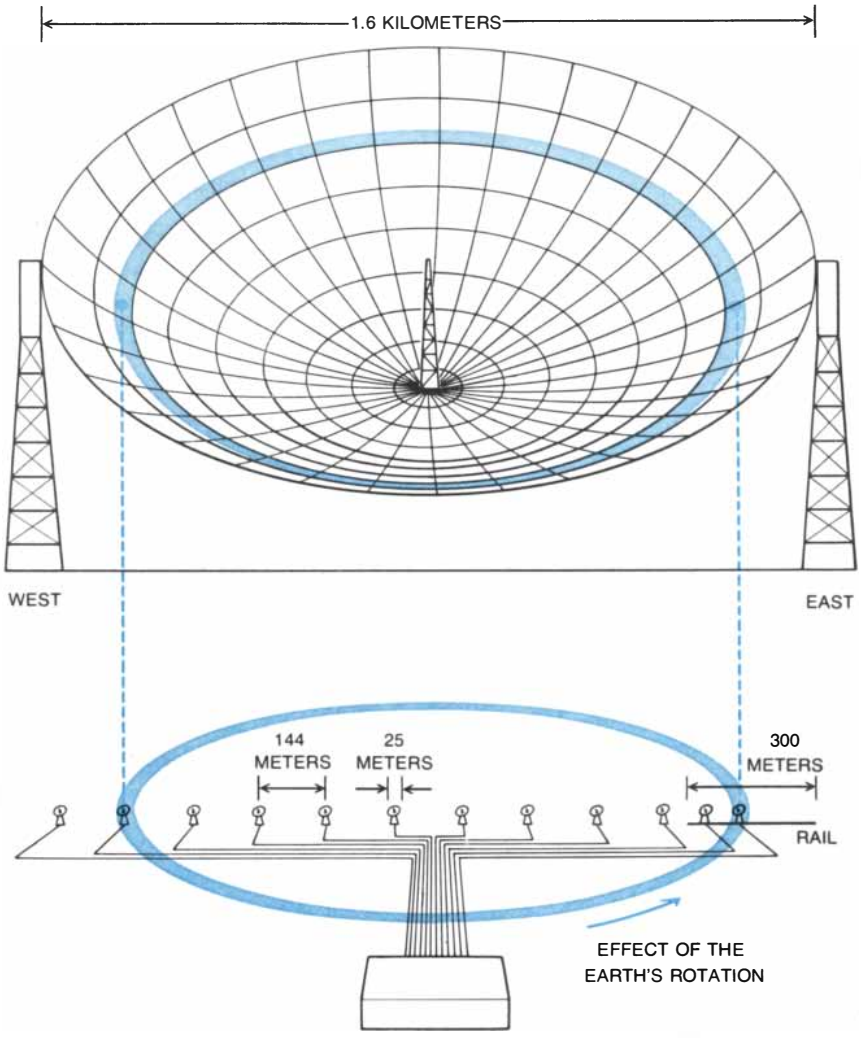
Since both sensitivity and angular resolution can be improved by building larger telescopes, much effort and money have been rightly spent on huge radio reflectors. The fact remains that the cost

of a single dish increases very rapidly with its aperture. For the telescopes needed to successfully study the structure of both radio galaxies and the universe, the engineering problems become insurmountable. An elegant technique first applied by W. N. Christiansen at the University of Sydney can simulate the response of an extremely large reflector with small reflectors. Martin Ryle of Cambridge became the chief advocate of employing interferometers with variable spacing to build up, or synthesize, apertures of a mile or more. His research culminated in a series of instruments that yielded high-resolution maps of radio objects. The latest Cambridge aperture-synthesis telescope, which spans a distance of three miles, has produced such maps of many sources, including Cygnus A.

Cygnus A is often considered the prototype of strong radio galaxies, and we may take its morphological features as being fairly typical of double sources. Most of the radio emission emanates from two compact regions near the outer edge of each component. Weaker emission, strung out behind the intense outer edges, almost bridges the gap to the optical galaxy located in the middle of the source. Energy released in the nucleus of the galaxy is somehow channeled to the two outer components, where its presence is made manifest by the intense radio emission. Coincident with the optical galaxy itself, as if to remind us of the violent events within it, is a third radio component, glowing like the hot embers of a quiescent fire.

The outer components of Cygnus A, expelled from the parent galaxy like two bullets from a double-ended gun, now span a distance of half a million light-years. How far can we expect them to travel before they finally stop, or cease to emit radio waves? The answer to this question depends on the energy available to drive the components and maintain their radio emission, the efficiency with which the energy can be transported to the components and the amount of material in the surrounding intergalactic space to resist the components' outward motion. Recently Tony Willis, Andrew Wilson and one of us (Strom) used the Westerbork synthesis radio telescope to search for abnormally large radio galaxies. We observed regions where several weak radio objects are known to lie unusually close to a strong source, in the hope of finding that they might all be part of a single huge object. This attempt to detect weak emission near an intense source is akin to someone's trying to discern a dimly lighted object while a strong light shines directly in his eyes. The great sensitivity and unrivaled stability afforded by the Westerbork telescope were both of paramount importance for such observations.

The first galaxy observed is designated DA 240. It has large, nearly circular components surrounding a central strong source [*see top illustration on page 27*]. Although in its bloated shape and in several other features DA 240 differs from Cygnus A, the two objects clearly belong to the same class. Both have a basic double structure and a compact central radio component. The central component of DA 240 helped in identifying the source with a distant visible galaxy of the elliptical type. The entire radio source spans an angular size of 34 minutes of arc in the sky; at the distance of DA 240 this corresponds to a linear dimension of just over six million light-years. The extent of DA 240 is thus more than 10 times greater than the diameter of Cygnus A and some 60 times greater



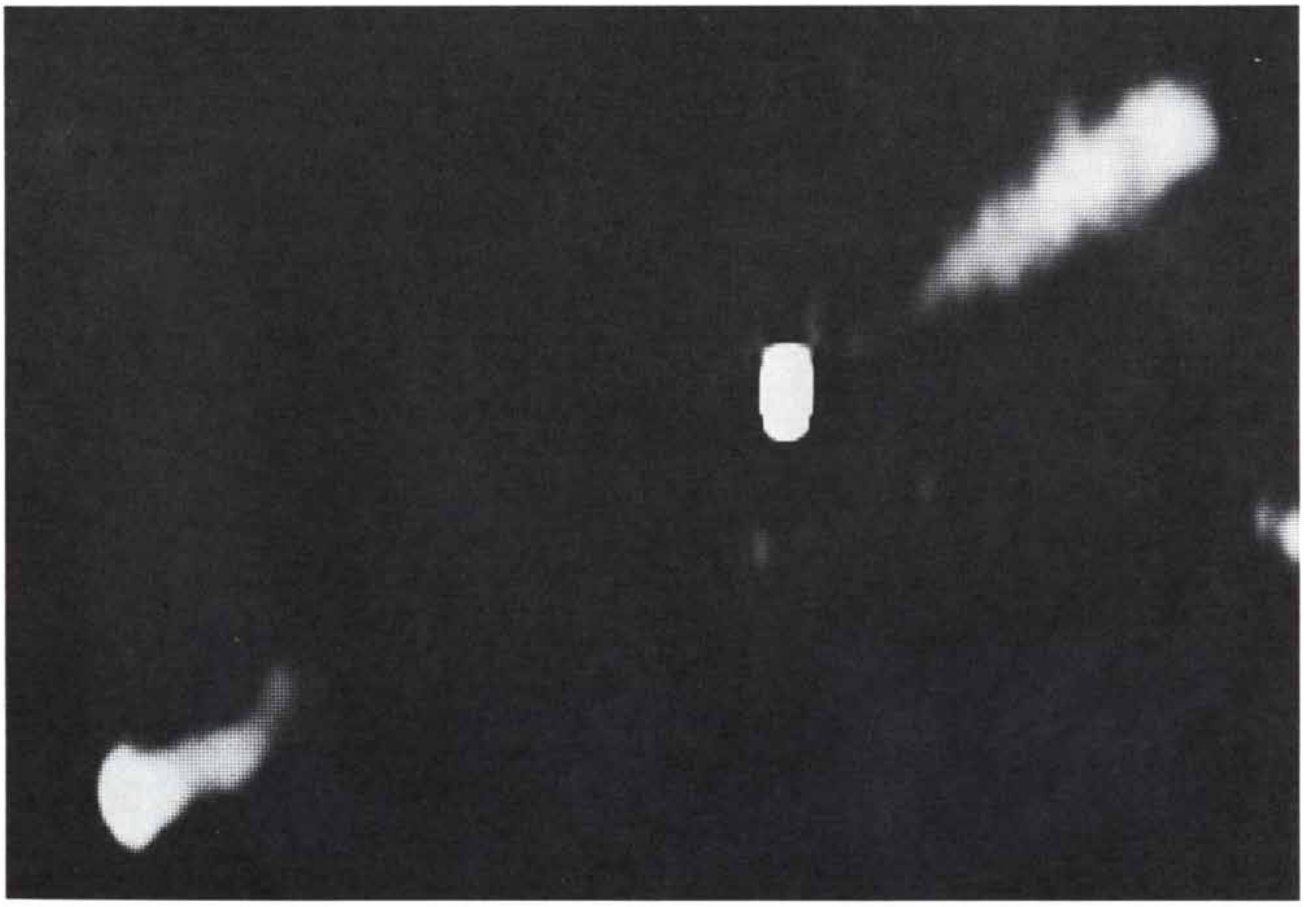
TECHNIQUE OF APERTURE SYNTHESIS is explained in terms of large and small radio reflectors. In a large dish (*top*) radiation is reflected from different elements, or small areas (*dark color*) of the surface, to the focal point of the telescope, where the signal is detected. In aperture synthesis the individual surface elements are essentially replaced with small reflectors (*bottom*). Instead of having a direct path to a common focal point the signals are carried by cables to a central location, where they are combined in pairs. At any one instant each pair of reflectors observes a celestial object at only one position angle. As time progresses, however, the earth rotates, causing each pair of reflectors to trace an annulus, or ring, under the source, changing the aspect of the source as it is seen from the array. Therefore in the course of a day the object is observed at all position angles and the annulus is completely filled in (*light color*). The size of the annulus depends on the separation between the pair of reflectors. When the observations from all the annuluses are combined, they have a resolution equivalent to that of a single large dish. At the Westerbork telescope the two movable dishes make it possible to conduct observations on four separate days with as many as 80 base lines ranging in length from 36 to 1,458 meters.

CONTOUR MAP OF CYGNUS A, a strong radio source in the constellation Cygnus, was made at a wavelength of six centimeters by P. J. Hargrave and Martin Ryle with an aperture-synthesis telescope at the University of Cambridge. There is an intense hot spot near the far edge of each of the two outer components, and weaker emission extends toward the central component. Measurements at higher resolution show that central component, which is associated with a visible galaxy, is extended in same direction as outer ones.

than the diameter of our own galaxy. An interesting feature of DA 240 is the circular appearance of the outer components. A similar feature is observed in several other large sources, but any thought we had entertained that it might be a general feature of such objects was quashed by the second galaxy we observed: 3C 236.

This object, a typical radio source associated with an elliptical galaxy, was discovered some 15 years ago. Subsequent studies made by Peter N. Wilkinson at Jodrell Bank showed that the radio emission originates from two compact components separated by one second of arc, equivalent to a physical separation of 6,000 light-years. On the map made with the Westerbork telescope the central double source lies between two extended components whose existence was unsuspected for more than a decade. Their presence means that the true angular size of 3C 236 is 39 minutes of arc, which is equivalent to nearly 18 million light-years.

The object held yet another surprise for us. Although the inner components and the outer ones differ in size by a factor of 3,000, the orientation of these structures in space is identical. Thus a line passing through the strongest por-



LARGEST OBJECT KNOWN in the universe is the radio galaxy designated 3C 236. The object is 18 million light-years from end to end and encompasses a volume larger than that of most clusters of galaxies. Radio photograph was made with Westerbork telescope at a wavelength of 49 centimeters. As with Cygnus A and DA 240, the central component of 3C 236 is associated with a visible galaxy.

tions of the two inner components would bisect each of the outer ones. This suggests that whatever initiates the double structure retains its sense of orientation for nearly the lifetime of the outer pair of components.

What might that lifetime be? By far the largest radio source known, 3C 236, could be the oldest radio source as well. From observations of other radio galaxies it has been estimated that radio components move away from the parent galaxy at speeds of a few thousand kilometers per second. Combining this average speed with the size of 3C 236 suggests that the object is about a billion years old.

Just as the longest bridge span is a stern test for civil engineering, so the oldest radio source will stretch our models and theories to the limit, tearing apart those that prove to be inadequate. Unfortunately detailed explanations of the development of double radio sources are in a far from satisfactory state, and the existence of such giant sources as DA 240 and 3C 236 only worsens an already critical situation. The difficulties can be highlighted by three questions: Why are two radio components expelled in opposite directions from some galaxies? How are the components confined and prevented from breaking apart? How is energy supplied to the components?

Let us consider the questions in reverse order. The radio emission we observe is generated at the expense of the energy carried by the fast subatomic particles. According to the precepts of synchrotron-radiation theory, a significant portion of that energy should be radiated away long before a radio source has had a chance to become as large as objects even considerably smaller than either DA 240 or 3C 236. Therefore there must be some mechanism that during the lifetime of a typical radio source either injects new high-energy particles into the components or reaccelerates the existing ones. If there is such a mechanism, energy must somehow be stored and gradually released as the radio components evolve.

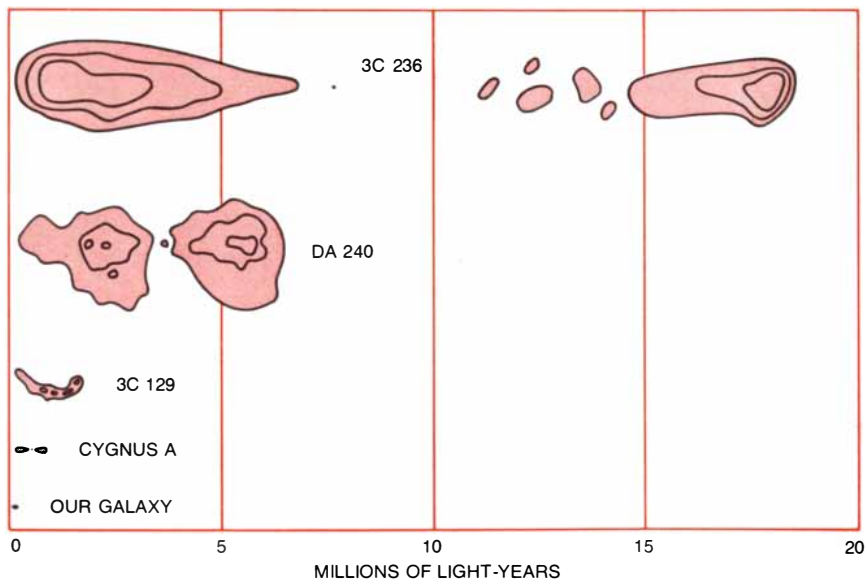
There is a class of theoretical models that assumes that every component carries its own supply of energy. The energy might be stored, for example, in turbulent, irregular motions of the hot gas permeating each radio source. As the fast-moving eddies collide with one another they gradually lose their kinetic energy; that energy in turn whips particles up to speeds sufficient to give rise to the synchrotron emission. In such a model the radio component is no more substantial than a cloud whose constituents—the hot gas, the fast particles and the magnetic field—all exert pressure and tend to cause the component to expand. We must then seek an explanation for the fact that most radio components have the confined appearance of those in Cygnus A and 3C 236 rather than the bulbous shape found in DA 240.

One potential confining agent is a tenuous intergalactic medium that resists the motion of the gas cloud moving through it. The cloud will be subjected to forces along its leading edge and sides that will tend to compress it and thereby hold it together. This ram pressure also decelerates the component. As the component slows down it encounters less resistance, the confining forces lessen and it expands. Therefore well-confined sources such as Cygnus A will eventually grow in size and evolve into very extended objects such as DA 240 [*see illustration on page 33*].

A serious flaw in this picture is that many radio sources, such as Cygnus A, have compact intense regions near the outer edge of their components. The volume enclosed by these hot spots is so small and the energy is so great that their containment by ram pressure alone appears unlikely. A second class of models assumes that the intense hot spots mark the presence of massive objects that are held together by their own gravitational attraction. As the objects plow into the intergalactic medium they gradually release energy that has been stored most probably in their rotational motion, and they leave behind a trail of radio emission. So far, however, none of the models proposing that the parent galaxy expels massive objects can explain the remarkable alignment of the inner and outer pairs of radio components found in 3C 236 and several other radio galaxies.

How can the nucleus "remember" the direction of the outer components for hundreds of millions of years? The answer is almost certainly that the components are expelled along the axis of rotation of a massive spinning body. Although such an axis can maintain a unique orientation for long periods of time, there is no obvious mechanism for expelling massive objects along it.

A third type of model, strongly advocated by Martin J. Rees of Cambridge, assumes that the nucleus of a radio galaxy is a storehouse that is continuously feeding energy to the hot spots in the outer components. The energy is channeled very efficiently in collimated beams of waves or particles moving at speeds close to the speed of light. Various ways in which two oppositely directed beams could be generated in galactic nuclei have been suggested. The beams could be produced by spinning massive objects or by high-pressure nozzles formed by rotating gas. Since the nozzles would be directed along the axis of rotation of the gas, the alignment of inner and outer pairs of radio components could be readily explained. Serious difficulties exist, however, in explaining



FOUR RADIO GALAXIES are compared in size with our galaxy, which is 100,000 light-years in diameter. They range from Cygnus A, 600,000 light-years across, to DA 240 and 3C 236, which are exceptionally large. Typical double galaxies are about size of Cygnus A.

how these beams remain stable over such vast distances and great spans of time.

Yet these problems beg the most intriguing question of all: Where does the energy come from? Just two entities—high-energy charged particles and the magnetic field through which they move—give rise to the synchrotron radiation that we observe. If the strength of the magnetic field were known, the intensity of the radiation would enable us at once to unambiguously compute both the number of particles producing the radiation and their average energy. The product gives the total energy of the particles in the source. The only unknown in this simple arithmetic is the strength of the magnetic field. Certain observations provide a rough estimate of its strength, but even without such information we can compute the minimum energy that must be contained in the radio components in order to account for the observed synchrotron emission. The energy thus calculated is tremendous; in the largest sources it is on the order of what would be released if all the hydrogen in a billion suns were fused into helium by thermonuclear reactions.

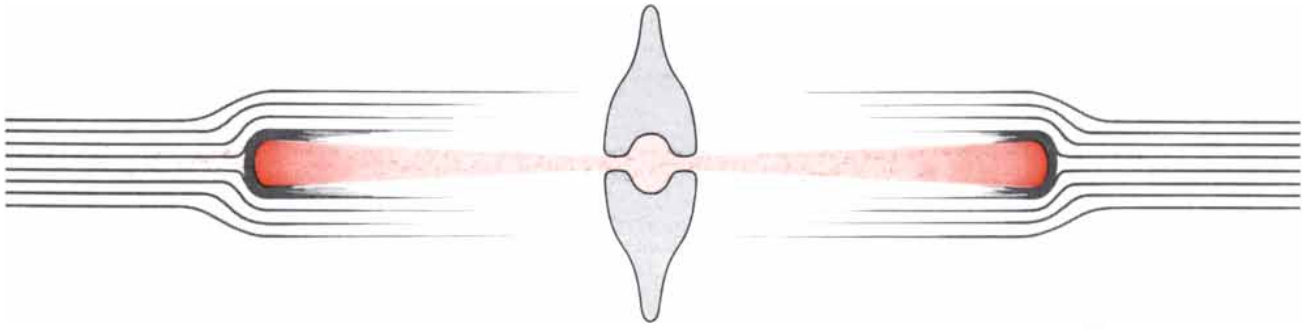
There are many indications that the ultimate source of the energy is situated in a very small region within the galactic nucleus, where it is improbable that such a large mass would be concentrated. Such considerations have led to the suggestion that a source of energy more powerful than fusion must be at work. There is only one such source that can be imagined within the known framework of physics: the total annihilation of matter. For a given mass such annihilation would yield 100 times more energy than fusion. Recent theoretical work shows that a large fraction of matter can indeed be converted into energy if the matter is compressed by the force of gravity to an extremely high density such as would be found in a black hole.

Is there any evidence that matter of such density exists? There are galaxies that have a very compact nucleus, such as the quasars and those that belong to the class known as Seyfert galaxies. These compact nuclei are the scene of explosive events of almost unimaginable magnitude that give rise to the same kind of high-energy particles that populate the components of the radio galaxies. Nothing is known about the cause of these events, but it is tempting to think that similar ones in radio galaxies have a role in the production and maintenance of the large double components. Actually what makes radio galaxies and their nuclei particularly fascinating is the expectation that in observing these phenomena we are essentially studying aspects of the great unknown territory of superdense galactic nuclei and possibly even black holes.

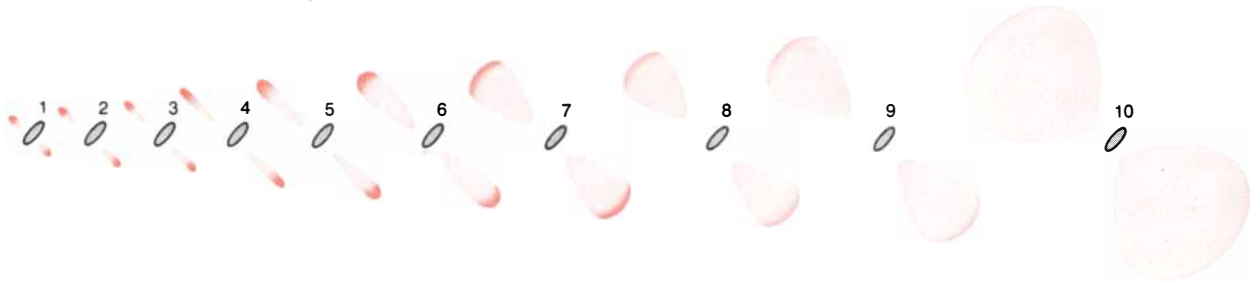
Some 30 percent of all extragalactic radio sources whose structure is known have a more complicated and irregular appearance than the symmetrical double sources we have been describing so far. Just as in biology the study of mutations enriches our knowledge of the evolution of normal species, so do recent observations of some of these deformed radio galaxies tell us a great deal about how the normal double sources evolve.

One giant galaxy that is now known to be relatively common is the head-tail radio galaxy, so named because of its tadpole-like appearance [*see bottom illustration on page 27*]. A strong radio component (the head) coincides with the visible nucleus of a galaxy, and a narrow, gossamer radio tail stretches away, sometimes for a distance of two or three million light-years. Nearly a score of radio galaxies with this characteristic shape are now known. Most of them have been found in rich clusters of galaxies, indicating that such clusters are a fertile spawning ground for tadpole sources.

A decade ago Ryle and one of his graduate students, M. D. Windram, used the One Mile Telescope at Cambridge to study a cluster of galaxies in Perseus. The Perseus cluster is one of the most densely populated, and it is relatively nearby at a distance of about 300 million light-years. Ryle and Windram showed that two bright galaxies at the edge of the cluster, NGC 1265 and IC 310, have clearly defined weak radio tails. They noticed that both of the tails point approximately away from the brightest member of the cluster, NGC 1275, a fact that seemed highly significant at the time. Measurements made in widely different regions of the electromagnetic spectrum indicated that NGC 1275 is one of the most violently exploding galaxies known. Clouds of hot gas and particles traveling at speeds close to the speed of light are being continuously expelled from its fireball nucleus. Ryle and Windram, developing an idea originally put forward by Geoffrey R. Burbidge of the University of California at San Diego, suggested that some of the hot gas is ejected with such tremendous force that it can travel millions of light-years across the cluster of galaxies. The remnants of the explosion from NGC 1275 might then collide with NGC 1265 and IC 310, "lighting up" radio sources



RAM PRESSURE is one mechanism suggested as the cause of the confined appearance of the gaseous components of double radio galaxies. If clouds of gas (*color*) are expelled at high speed into a tenuous intergalactic medium (*dark gray*), pressure builds up at the leading edge of the clouds that confine the gas. It also tends to concentrate the gas near the leading edge of the cloud, resulting in stronger radio emission from this outermost region. Ram pressure plays a part in several theoretical models. In the model shown here the gas subjected to the ram pressure originates near the center of the galaxy and is initially confined by a cooler surrounding cloud. As its internal pressure increases, the hot gas tends to tunnel through the confining cloud along its axis of rotation. Eventually two nozzles form, each producing a collimated beam of waves and particles rather like rocket exhaust. The waves and particles in the narrow beam interact very little with the intergalactic medium. Thus energy from the center of the galaxy can be transported almost without loss to the two hot spots in the radio components. As source evolves, beam pushes hot spots farther from parent galaxy.



MORPHOLOGICAL DEVELOPMENT such as DA 240 might be explained by the mechanism of ram pressure. When the double radio components are first ejected, they are extremely compact (*1 through 5*) and the radio emission is strongest from the outermost regions. The ram pressure from the intergalactic medium then decelerates the components; as a result of their slowing down, pressure decreases and components begin to expand laterally (*6 through 8*). With drop in gas's density from expansion, the radio emission becomes less intense. Eventually (*9 and 10*) huge diffuse components merge with intergalactic medium and emission disappears.

and spinning out the tails as they blow past.

There were always several inherent difficulties with this picture of interacting galaxies, and a large body of observational evidence has recently built up against it. One objection is that for the dozen other head-tail sources that have since been discovered in other clusters there are no obviously peculiar galaxies nearby that could be held responsible for igniting the radio emission. A second objection comes from new observations of the Perseus cluster. Because only a few of the thousands of galaxies in the cluster are detectable radio sources, the hot gas emitted from NGC 1275 would have to travel in highly directional streams. Therefore the radio tails should all point directly away from NGC 1275. Three years ago a weak tadpole radio galaxy was found, however, whose tail pointed roughly not away from NGC 1275 but toward it. The question was finally answered when high-resolution measurements showed that there are compact radio components in the nucleus of nearly all head-tail galaxies. This means that the emission is almost certainly caused by explosive events within the nucleus of the galaxies themselves and not by some interaction with an external source such as NGC 1275.

The compact radio components are like the ones we see so often in the symmetrical double radio sources. Moreover, the head-tail galaxies resemble the simple double sources in several other ways. For example, in some of them one can see that both the head and the tail have a pronounced double appearance [*see illustration on next page*]. Because of such similarities it seems more logical to seek an explanation for the head-tail sources within the general framework of double radio galaxies.

Three years ago Cesare Perola of the University of Milan and Piet van der Kruit, Harry van der Laan and one of us (Miley) at the Leiden Observatory suggested that head-tail galaxies are active in their own right and that the observed tail is due to a trail left behind as the galaxies move through the cluster. From measurements of the Doppler shifts in the visible lines in their spectra we know that galaxies in clusters are indeed moving with respect to one another, often at speeds of a few thousand kilometers per second. For instance, the tailed radio galaxy NGC 1265 has a velocity along the line of sight of some 2,300 kilometers per second more than the mean speed of the rest of the galaxies in the Perseus cluster. If we assume that the speed of the galaxy perpendicular to the line of sight is not too different from its line-of-sight velocity, NGC 1265 would need about 100 million years to travel the length of its radio tail. That span of time is roughly comparable to the age of the radio source itself, which we can deduce independently from its radio spectrum and the theory of synchrotron emission. The agreement between the two time scales adds weight to our picture of a galaxy trailing energetic particles behind it as it travels through space. Therefore the electrons we now see at the end of the tail were probably deposited in the surrounding medium 100 million years ago by the passing nucleus of NGC 1265.

If the radio tail indeed defines the galaxy's path through the cluster, we can identify each point on the tail with the position of the galaxy at a known time in the past. Just as the earth's sedimentary layers are a record of the geological past, so has the galaxy written its radio history on the sky. What can this unique record tell us?

A high-resolution radio contour map of the head of NGC 1265 shows that there are discrete blobs of radio emission. This suggests that the radiating particles are ejected not smoothly and continuously from the nucleus but in a series of bursts. These bursts are expelled like puffs of smoke from a steam locomotive. In each puff two clouds of high-energy particles are ejected in opposite directions, forming a miniature symmetrical double radio source.

The basic mechanism that gives rise to the miniature double source is probably the same one that gives rise to the large symmetrical double sources. In the case of the head-tail galaxies, however, the fast motion of the galaxy prevents the buildup of a large double source. The gas clouds are expelled with a speed comparable to the speed through space of the galaxy itself, and they are slowed by ram pressure from the surrounding gas. Eventually they are stopped, and adjacent pairs of blobs merge to form a smooth double radio tail. From the distance between successive pairs of radio blobs we can deduce that such outbursts in the galactic nucleus recur every few million years.

The simple picture of a puffing radio galaxy hurtling through the cluster has been developed further at the Leiden Observatory by Walter Jaffe and Perola. They have suggested that a magnetic field, similar to the field of a bar magnet, is attached to the galaxy. As the galaxy plows through the gas within the cluster the galactic magnetosphere is dragged behind it, rather like the geomagnetic tail of the earth in the wind of charged particles from the sun. Explosive puffs from the galactic nucleus then inject energetic electrons into the preexisting magnetic field, and the electrons radiate by the synchrotron process.

One of the attractive features of the model is that it makes some clearly observable predictions. The most obvious is that the magnetic field should be

aligned along the radio tail. Recent measurements of the NGC 1265 radio source with the instrument at Westerbork by Kelvin Wellington, van der Laan and one of us (Miley) show that the magnetic field is highly regular and is indeed directed along the tail. Although other explanations are still possible, the weight of the observational evidence available at present supports this picture.

It should now be clear why the head-tail sources not only provide an opportunity to study the evolution of radio sources but also are a useful probe of the physical environment within clusters of galaxies. In particular the general appearance of the head and the tail provides one of the few strong arguments in favor of the presence of a medium between the galaxies. For example, the fact that the long, narrow tail of NGC 1265 has an almost constant width is strong evidence that each pair of gas blobs is slowed down and stopped soon after it has been ejected from the nucleus of the galaxy.

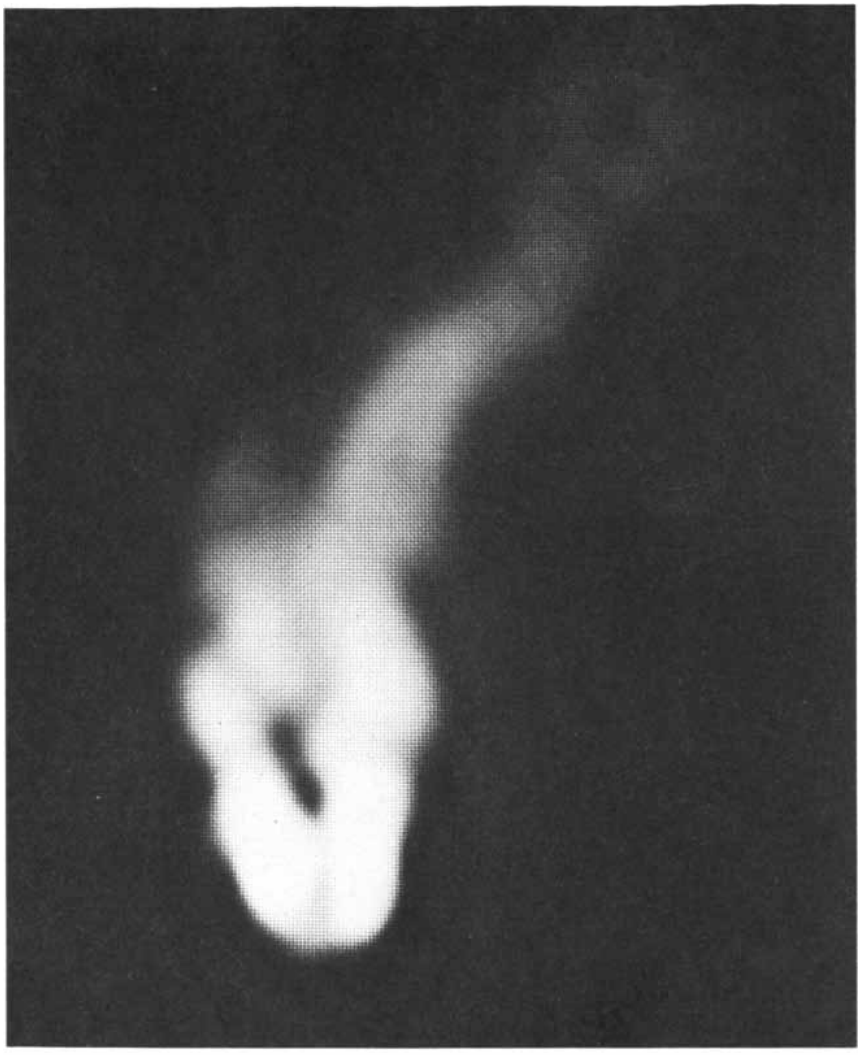
What is responsible for this retardation? So far the only agent suggested for decelerating the blobs is the pressure of a gas surrounding the galaxy. One can estimate that the gas must have a density roughly equivalent to three hydrogen atoms per liter. Independent evidence for the existence of this tenuous gas comes from X-ray astronomy. Several nearby clusters of galaxies, including the Perseus cluster, have been observed to emit X rays. The most likely source of the X-ray emission is a very hot low-density gas permeating the cluster. Such a hot gas, with a temperature of about 100 million degrees K., could both produce the observed X-ray intensity and confine the radio tails.

A further hint regarding the conditions in the intergalactic medium is provided by the fact that many of the tails curve. One explanation for the curvature is that the exploding galaxy follows a curved orbit in the gravitational field of the cluster. An alternative explanation is that the bending is brought about by large-scale turbulence in the intergalactic medium. The medium would then distort the tail just as the smoke trail from a steam locomotive is deformed on a windy day. Future studies of the curvature of the tails and the mechanisms that produce it promise to reveal still more about the conditions inside clusters of galaxies.

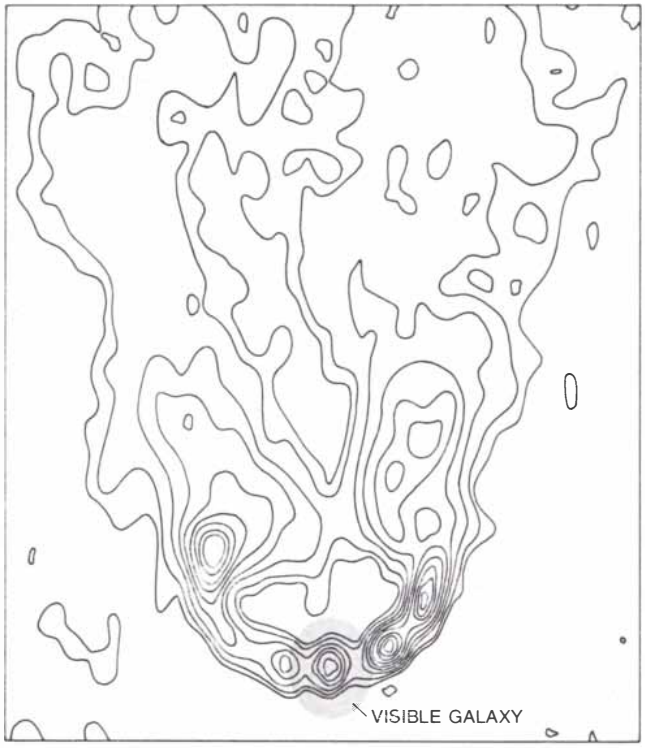
Several other basic problems regarding the head-tail radio sources remain unsolved. For example, in spite of the striking similarities between such sources and normal double radio galaxies, the head-tail sources tend to be intrinsically weaker in their radio emission. Could the motion of the parent galaxy alone impede the formation of more luminous radio sources? Moreover, why do head-tail sources exist in some clusters such as the Perseus cluster and not in others? Does their presence indicate that the intergalactic gas is denser (or hotter) or simply that the galaxies move at greater speeds?

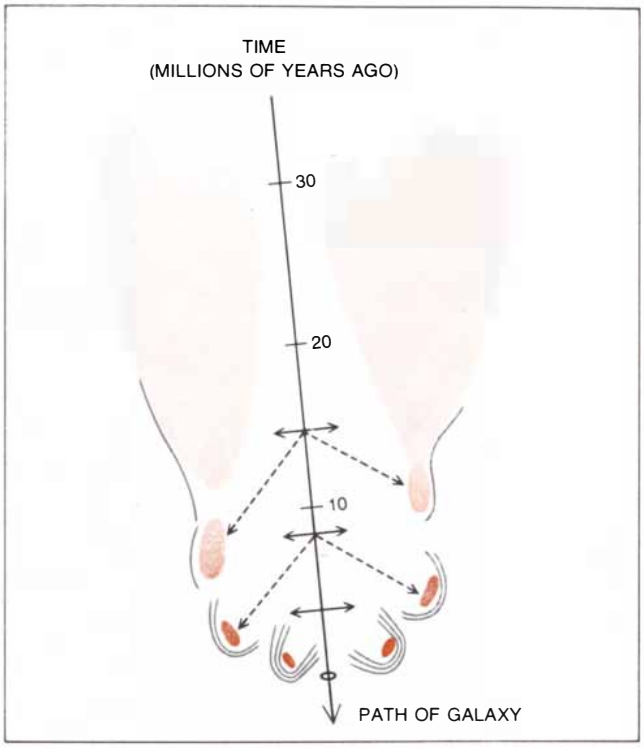
The period since the birth of radio astronomy more than 40 years ago has had a turbulent history of unexpected discovery. Some of the most exciting developments have been connected with radio galaxies. The detailed shapes of the radio sources have for the first time given direct proof that a tenuous gas fills the space between the galaxies. Among the leading unsolved problems of astrophysics, however, are the cause of the characteristic double structure of such objects and the source of their vast energy, which may require the invocation of some novel process such as the total annihilation of matter.

The study of the double structure of radio galaxies clearly extends into the unknown territory of the dense galactic nuclei, for whose exploration radio interferometers extending almost the full diameter of the earth are now combining forces with the largest optical telescopes. The enterprise has also been joined by X-ray observers, who work at wavelengths from 1,000 to 10,000 times shorter than those of light. Improved radio synthesis arrays, the increased use of intercontinental interferometry and more sensitive X-ray detectors will undoubtedly bring closer the solution of the intriguing problems presented by the giant radio galaxies.

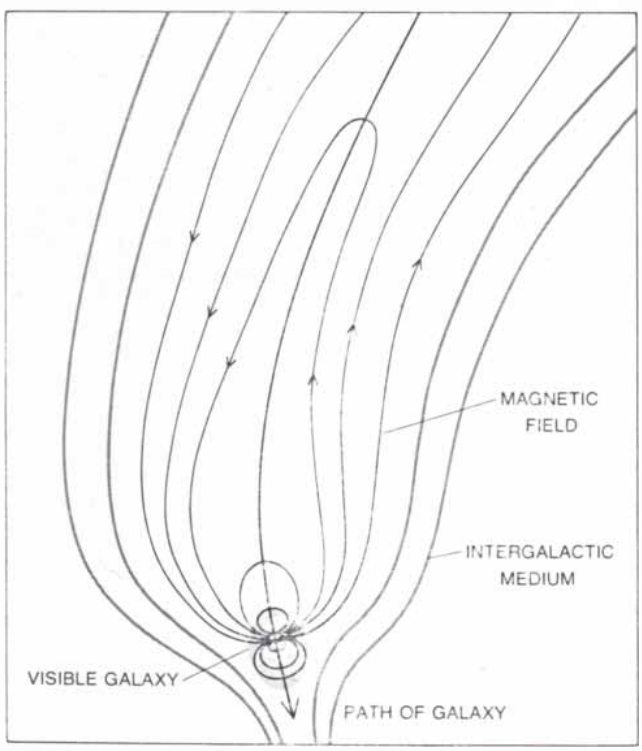


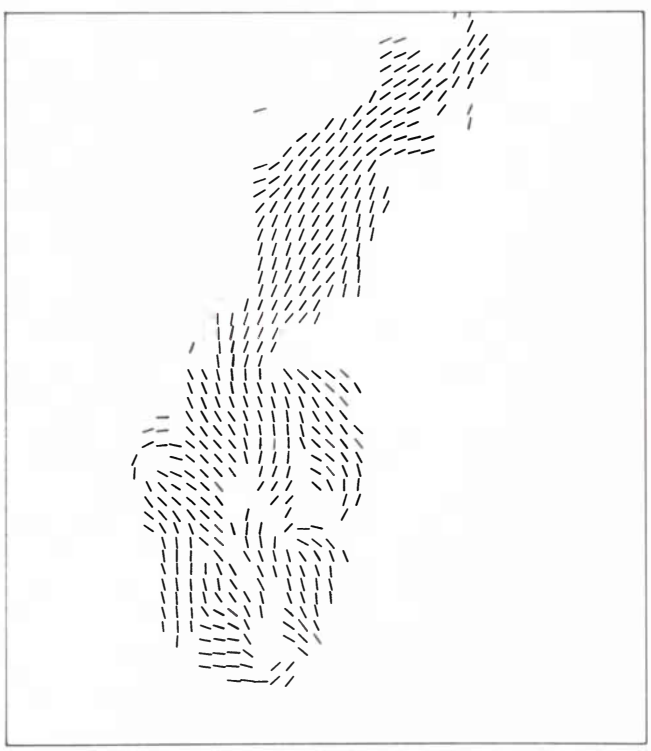
HEAD-TAIL RADIO GALAXY NGC 1265 in the Perseus cluster of galaxies is shown in a radio photograph that was made with the Westerbork telescope at a wavelength of 21 centimeters. Both the head and the tail of the object appear to have a double structure.





PUFFING RADIO GALAXY hurtling through intergalactic space is the model that has been proposed to account for the observed structure of NGC 1265. The radio contour map of the head of NGC 1265 (*left*), made with the Westerbork telescope at a wavelength of six centimeters by Kelvin Wellington, Harry van der Laan and one of the authors (Miley), has a higher definition than the radio photograph on the opposite page. According to the model, as the galaxy moves through space it occasionally releases twin puffs of energetic gas (*right*). The gas is slowed by the intergalactic medium and eventually trails behind. From the velocity of NGC 1265 it is possible to estimate when puffs were emitted. Any point on the galaxy's line of travel therefore corresponds to a time in its history.





MAGNETOSPHERIC MODEL of NGC 1265 suggests that the galaxy has a dipole magnetic field that is dragged behind it (*left*) as the galaxy travels through the cluster. High-energy particles from the exploding nucleus of the galaxy are injected as bursts into the **magnetic field, where they "glow" with synchrotron radiation. Observations of the direction of the magnetic fields in the gas of NGC 1265 have shown that, as is predicted by the magnetospheric model, the magnetic field of galaxy is indeed aligned along its tail (*right*).**

HOW BACTERIA SWIM

The helical filaments of the thin flagella that propel bacteria do not wave or beat but instead rotate rigidly like propellers! And they are driven by a reversible rotary motor at their base

by Howard C. Berg

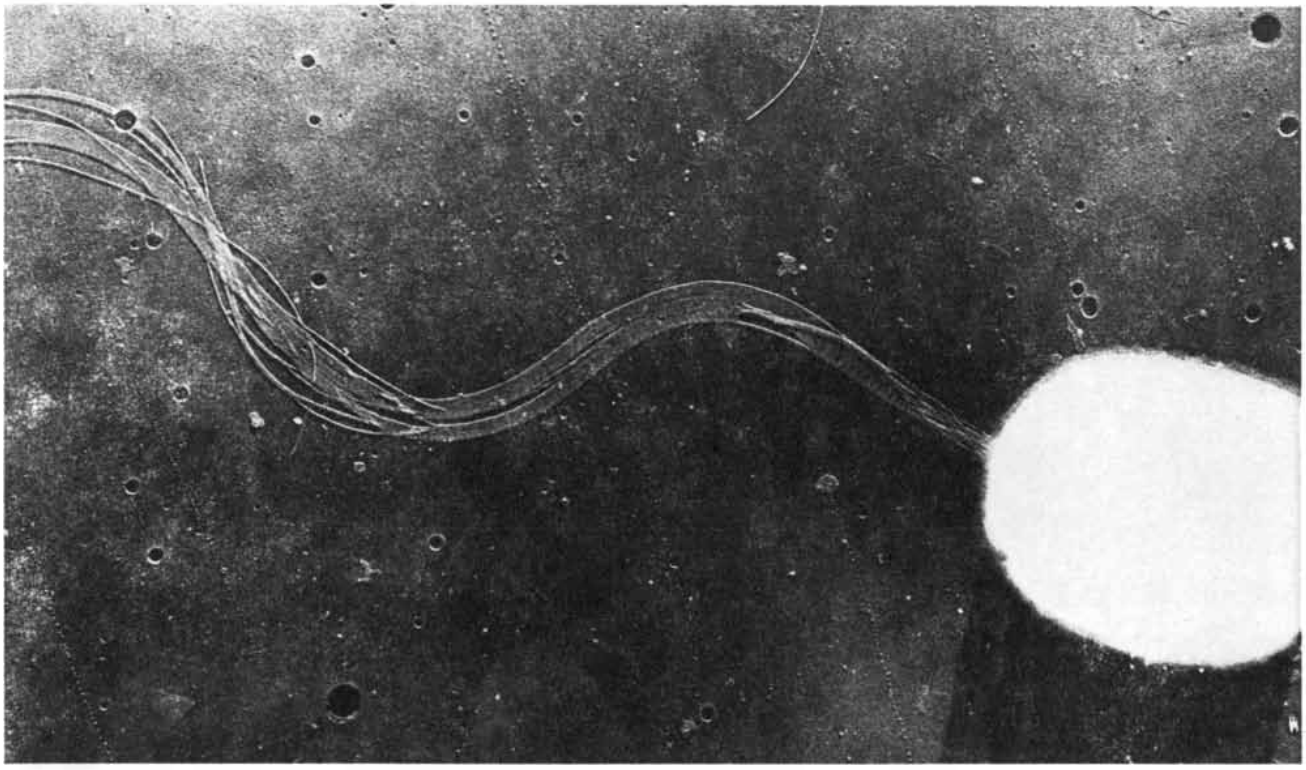
When Anton van Leeuwenhoek looked through a single-lens microscope in 1676 and obtained man's first recorded glimpse of bacteria, it was their motion that most delighted him: "I must say, for my part, that no more pleasant sight has ever yet come before my eye than these many thousands of living creatures, seen all alive in a little drop of water, moving among one another, each several creature having its own proper motion." Leeuwenhoek's observations marked the beginning of microbiology, yet the motion he described is only now beginning to be understood in detail. Bacteria swim by moving their flagella: thin, whiplike structures that arise at one point or more on the cell surface. Each flagellum has three parts: a helical filament, a short segment called the hook and a basal structure embedded in the cell wall and cytoplasmic membrane. The filament has long been thought to propagate helical waves, to propel the cell by undergoing rhythmic changes in shape. It now appears that it rotates rigidly. A bacterial flagellum is not a bending machine; it is a propeller, and it is driven by a rotary motor.

It was the German microbiologist C. G. Ehrenberg who first recognized bacterial flagella as organelles of locomotion. In 1838 he described *Chromatium okenii,* a large "purple" bacterium destined to play an important role in early work on bacterial motility. It is found in water containing hydrogen sulfide or other reduced sulfur compounds, which it utilizes with carbon dioxide in a form of photosynthesis. It is propelled by a bundle of some 40 flagellar filaments that arise at one pole. Its movements were described in 1883 by the German physiologist T. W. Engelmann, who was attracted by its peculiar behavior in response to changes in light intensity. When Engelmann observed the bacteria through the microscope for some time under ordinary lighting conditions, they swam steadily across the field of view. If the intensity of the light was decreased suddenly, all the cells rushed back several body lengths, stopped and then resumed their normal motion. Engelmann called this maneuver a shock reaction because it gave him the impression of fright. He found he could elicit an identical response by suddenly exposing the water droplet containing the bacteria to carbon dioxide.

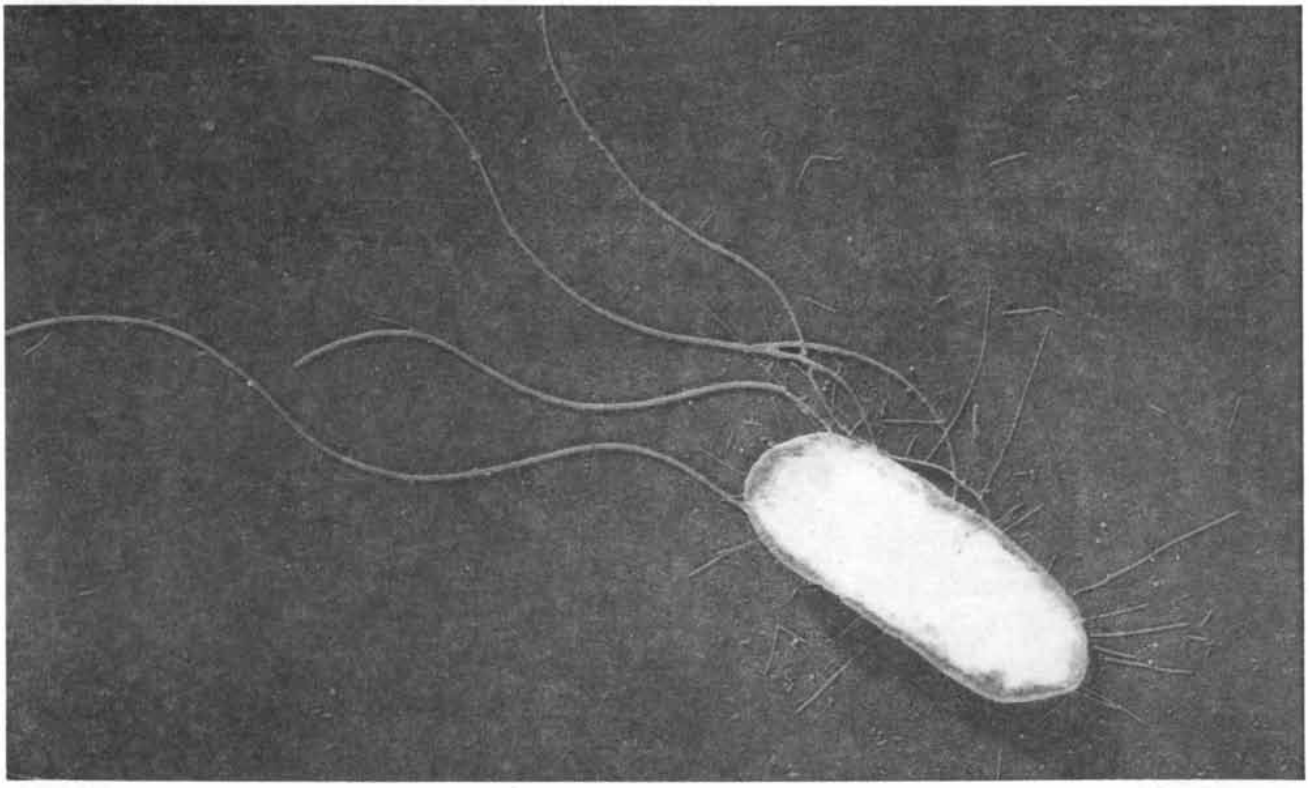
The shock reaction was described in more detail in 1915 by the German botanist Johannes Buder, who observed the cells in visible light as he stimulated them in the far red, a portion of the spectrum to which they are particularly sensitive. Under constant illumination the flagellar bundle appeared to rotate as a left-handed screw, pushing the cell [*see illustration on page 38*]; the body turned more slowly about the same axis, in the opposite direction. When Buder reduced the intensity of the stimulus light, the bundle slowed down, stopped (retaining its helical shape) and then started up in the opposite direction, now pulling the cell. If he applied a second stimulus, the bundle changed its direction of rotation again and the cell resumed its normal motion. If the second stimulus was not applied, a second reversal occurred spontaneously a few seconds after the first; Buder assumed that it was a mechanical event implicit in the workings of the flagella rather than a response to the original stimulus. When he photographed cells with an exposure of 1/25 second, the images of the bundles were completely blurred, implying that their rate of rotation exceeded 25 revolutions per second. A few years later the German microbiologist Peter Metzner examined the cells under stroboscopic lighting and found that the rotation rate (with respect to the cell body) had to be at least 40 to 60 revolutions per second. The body meanwhile rotated from six to eight times a second and moved forward at about a tenth of a millimeter per second.

Neither Buder nor Metzner imagined that the flagellar filaments actually rotated rigidly. They interpreted their results in terms of the theory of wave motion proposed in 1884 by the zoologist Otto Bütschli, who worked not with bacteria but with one-celled animals, protozoa with helical flagella. Bütschli assumed (correctly in the case of protozoa) that the flagellum was fixed to the cell body at its base and that it bent in a helical fashion. Motion of this kind can be demonstrated by threading a flexible tube over a rigid helical wire, holding one end of the tube fixed and rotating the wire [*see upper illustration on page 39*]. The rotation causes the tube to propagate a helical wave. The tube looks as if it were rotating, but it is not; it simply bends laterally. Bütschli suggested that such bending was generated by the contraction of parallel elements running the length of the flagellum, with the longest (the most relaxed) elements along the outer surface of the helix and the shortest (the most contracted) along the inner surface. Such a flagellum could propagate a helical wave if successive elements contracted in a rotary sequence; the direction of propagation would change if the direction of the sequence changed.

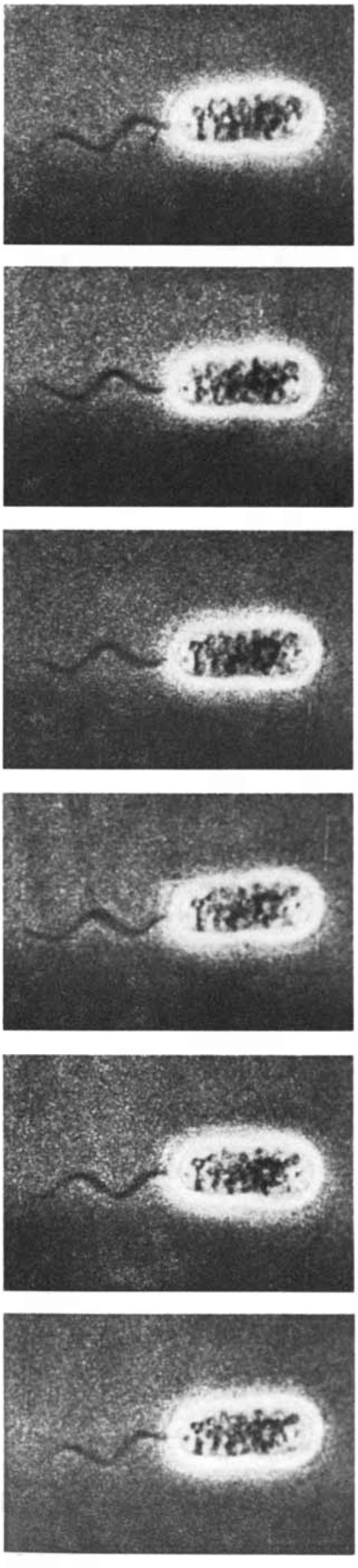
Bütschli realized that either rigid rotation or helical-wave propagation would generate both thrust and torque, so that the cell body would be both



CHROMATIUM OKENII, a large bacterium that was the subject of early studies of bacterial movement, is shadowed with platinum-palladium and enlarged 12,000 diameters in an electron micrograph that was made by H. Kutzner of the Technische Hochschule in Darmstadt. As many as 40 flagellar filaments arise at one pole of the bacterium. They form a bundle that either pushes or pulls the cell. Mature *C. okenii* are from eight to 15 micrometers (thousandths of a millimeter) long and the flagella are about 25 micrometers long.

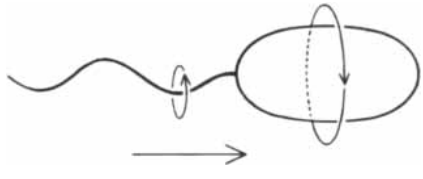


ESCHERICHIA COLI is shadowed with platinum and enlarged 26,000 diameters in an electron micrograph made by Ginny Fonte of the University of Colorado. Flagellar filaments (usually about six of them, some seven micrometers long) arise at random at the sides of *E. coli*. The filaments either form a flagellar bundle that pushes the cell or function independently, causing the cell to "twiddle," or jiggle about. In addition to the flagella there are thinner, straight fibers: Type I pili, whose function is not known.



pushed (or pulled) and made to rotate. That would be a natural consequence of the resistance of the medium to the motion of the flagellum [*see lower illustration on opposite page*]. The resistance is due to viscous forces; inertial effects are negligible. One implication is that it is impossible for a bacterium to coast; when the flagella stop moving, the cell comes to a standstill within about a millionth of a body length. (A man would have a similar experience if he tried to swim in asphalt.) The hydrodynamics of the motion of bacteria that, like *C. okenii,* have polar flagella are now well understood.

The current emphasis in work on bacterial motility is on the molecular biology of behavior. As was evident with *Chromatium,* bacteria not only swim but also modify the way they swim in response to changes in their environment. How do they detect and process sensory information? The intestinal bacterium *Escherichia coli* has proved to be the most suitable organism for such studies because so much is known about its genetics and biochemistry. It is much smaller than *Chromatium* and its flagellation is different: about six filaments arise at random sites on the sides of the cell and come together to form a bundle that emerges near one pole. In 1969 Julius Adler of the University of Wisconsin showed that *E. coli* cells have chemoreceptors: proteins sensitive to specific changes in the composition of the external medium that influence the cells' behavior. It had been known for some time that bacteria accumulate in regions that appear more favorable to them chemically, but the work on photosynthetic bacteria had led people to assume that such responses depended simply on differences in the amount of energy made available to the flagella. Adler found that *E. coli* could respond to substances it could not metabolize or even transport across its cytoplasmic membrane and that sensitivity to sets of structurally related compounds could be abolished by point mutations.



SWIMMING C. OKENII is shown in six successive frames from a film made at 600–1,000 frames per second by Norbert Pfennig of the University of Göttingen. In about a hundredth of a second the flagellar filaments complete one turn and the cell body moves forward a small fraction of its length. The drawing, adapted from one made by Johannes Buder in 1915, shows the direction of rotation of the filaments and the cell body.

I decided to try to make detailed measurements of the movements of *E. coli* and learn just how they respond to chemical stimuli. The bacterium is only about a micrometer (a thousandth of a millimeter) in diameter and two micrometers long, and it can swim some 20 diameters a second; at the magnifications at which it can be closely observed it swims out of focus in a fraction of a second. If one attempts to confine it between a slide and a cover slip, it sticks to the glass or swims in circles, something it does not otherwise do. I therefore undertook while at Harvard University to develop a microscope that automatically follows the movement of an individual cell in three dimensions [*see illustration on page 40*]. The instrument tracked its first bacterium in April, 1970, and has since then been moved to the University of Colorado.

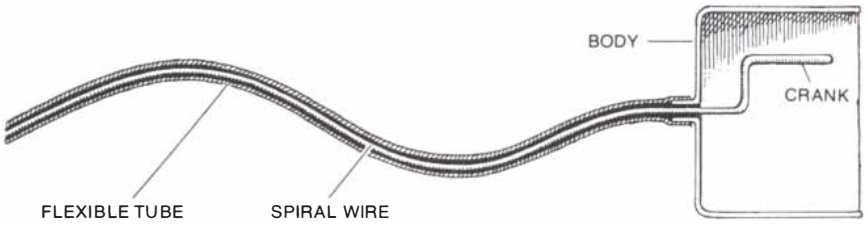
The bacteria are suspended in a small windowed chamber about two millimeters in diameter that is mounted on an electromechanical transducer. The image of the cell is focused at the ends of six optical fibers, each of which leads to a photomultiplier. The fibers are mounted in pairs. One pair looks at the right and left sides of the image, another at the front and back of the image and a third at the top and bottom of the image (above and below the plane of focus). The image is bright in a dark field. The differences in the light outputs of the two members of each pair are converted into error signals that are amplified to control the position of the transducer. The feedback is negative: the chamber is moved in the direction that reduces the amplitudes of the error signals, that is, in the direction that keeps the image centered on the fiber-optics assembly. The bacterium thus remains at a fixed point in the laboratory and the displacements of the transducer (which keep it there) provide a record of its movement through the medium in which it is suspended.

The *E. coli* cell executes a three-dimensional "random walk" [*see illustration on page 41*]. It "runs," or swims steadily along a gently curved path. Then it "twiddles," or stops and jiggles about. Then it runs again. A twiddle is a random event. It is just as likely to come at one instant of time as at anoth-

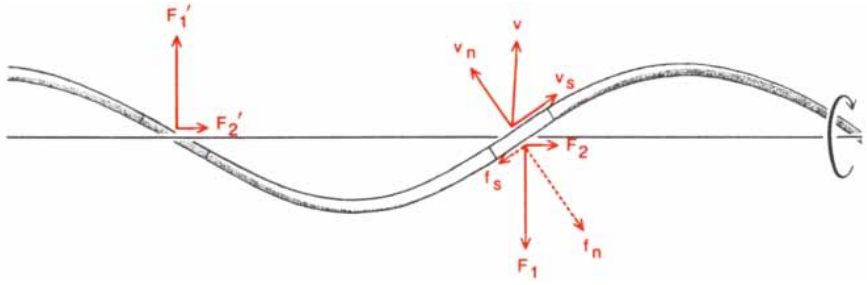
er; the probability of its occurrence is a constant. The probability that a twiddle will end is also a constant. Twiddles occur about once a second and last about a tenth of a second. The direction of the next run is chosen almost at random: in mediums of low viscosity there is a slight bias in favor of runs in the forward direction; in mediums of high viscosity the cells often back up. The change in direction is too abrupt to be explained by the Brownian motion of the bacterium, so that the flagella must remain active during a twiddle. The flagellar bundles of *E. coli* have been seen to push the cells but not to pull them. A twiddle apparently begins when the bundle flies apart; this has been observed in *Salmonella typhimurium*, a species closely related to *E. coli*, at the University of California at Berkeley by Robert M. Macnab, now at Yale University, and Daniel E. Koshland, Jr., who equipped their dark-field microscope with a high-intensity xenon arc. When the twiddle ends, the bundle forms again and the cell, now headed in a new direction, sets off on the next run.

If a chemical known to attract *E. coli* is added to the tracking chamber so that its concentration is higher on one side than on the other, the random walk becomes biased. Whenever the bacterium moves up the gradient (toward the region of higher concentration), the probability of a twiddle is smaller than it is in an isotropic, or uniform, solution; whenever the bacterium moves down the gradient, the probability stays about the same as in the isotropic solution. The result is that the bacterium drifts up the gradient by increasing the length of runs that take it in the favorable direction. Other parameters of the random walk remain constant. These findings were surprising in the light of earlier work, including the work with *C. okenii*, which had implied that bacteria respond only when conditions become less favorable.

A more quantitative measure of the stimulus-response relation can be obtained by generating or destroying an attractant with an enzyme; in this case the medium remains isotropic, but the concentration of the attractant changes over time. Such an experiment was performed with the tracking microscope by Douglas A. Brown, who is now at the University of California at Santa Barbara. When the concentration of the attractant increases, the cells twiddle less frequently. When the concentration decreases, the twiddles come about as often as they do in the absence of a stimulus. The data fit a theoretical model in which the probability of a twiddle depends on the time rate of increase in the number of attractant molecules that become bound to chemoreceptors.



BENDING AND ROTATION are difficult to distinguish in a helical flagellum, as is demonstrated by this model, which represents a nucleated cell with such a flagellum. A flexible tube is threaded over a rigid spiral wire and connected rigidly to the cell body; the wire ends in a crank. When the body is held fixed and the crank is turned, the tube propagates a helical wave. The tube seems to be rotating, but it is not; it is simply bending laterally.



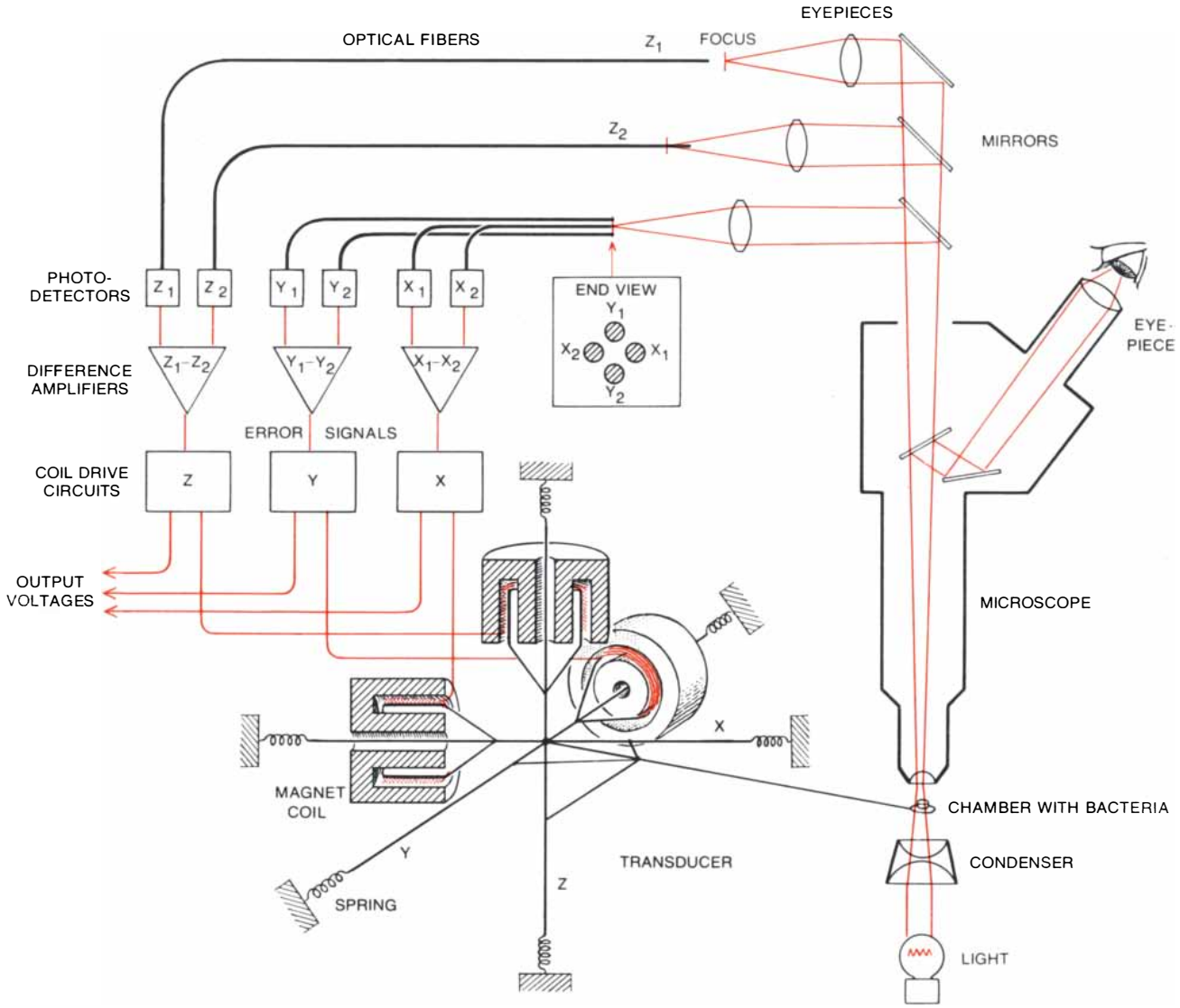
VISCOUS FORCES, originally worked out for a filament propagating a helical wave, apply also to a rotating filament. Forces perpendicular to the axis but acting in opposite directions tend to rotate the cell body; forces parallel to the axis tend to push (or pull) the cell body. The velocity (v) of a short segment of the helix is divided into a perpendicular component (v_n) and a parallel one (v_s). Viscous drag is proportional to the velocity, but is larger for motion perpendicular to the segment than for motion parallel to it. The components of the drag perpendicular to the segment (f_n) and parallel to the segment (f_s) are resolved into a component (F_1) that contributes to the torque and a component (F_2) that contributes to the thrust. The results of a similar analysis are shown for a segment (*left*) half a turn away along the helix. F_1 and F_1' are equal in magnitude but opposite in direction; F_2 and F_2' act in the same direction. The arrow at the right shows the direction of apparent (or real) rotation.

Another means of stimulating the cells was devised by Macnab and Koshland, who mixed bacteria (*S. typhimurium*) suspended in a solution containing an attractant at one concentration with a second solution that contained the attractant at a different concentration. When by this means the concentration of an attractant is increased by a large amount, the cells swim for several minutes without twiddling and then gradually resume their normal mode of behavior. When the concentration of the attractant is decreased by a large amount, the cells twiddle more frequently, but only for a few seconds. Macnab and Koshland found that chemicals known to repel *E. coli* have the opposite effect: twiddles are suppressed dramatically when the concentration of a repellent is decreased.

It was in the course of tracking bacteria that I became fascinated with the question of how the flagella work. The filaments are extraordinarily thin, only about a hundredth of a micrometer in diameter (or about half the diameter of the cell organelles known as microtubules). They are polymers of flagellin, a protein with no known enzymatic activity. Their protein nature was established in 1948 by Claes Weibull, who is now at the University of Lund, during a spirited debate with Adrianus Pijper of the University of Pretoria; Pijper maintained that what looked like filaments were merely "polysaccharide twirls" emitted by the bacteria as artifacts of locomotion. The morphological and chemical properties of the filaments have since been studied by a number of workers, including Henry Koffler of Purdue University, Tetsuo Iino of the University of Tokyo, Sho Asakura of Nagoya

University and Goro Eguchi of Kyoto University. The shape of the filaments varies depending on the amino acid composition of the flagellin and on environmental conditions such as temperature, acidity and dynamic load. When the filaments are suspended in a physiological salt solution and heated, they dissociate. If such a solution is cooled and then seeded with fragments obtained by breaking up intact filaments with sound waves, new filaments form; the crystallization always begins at one end of the seed. In live bacteria growth takes place at the tip of the filament, the molecules of flagellin moving out to the site of crystallization through the core of the filament from its base.

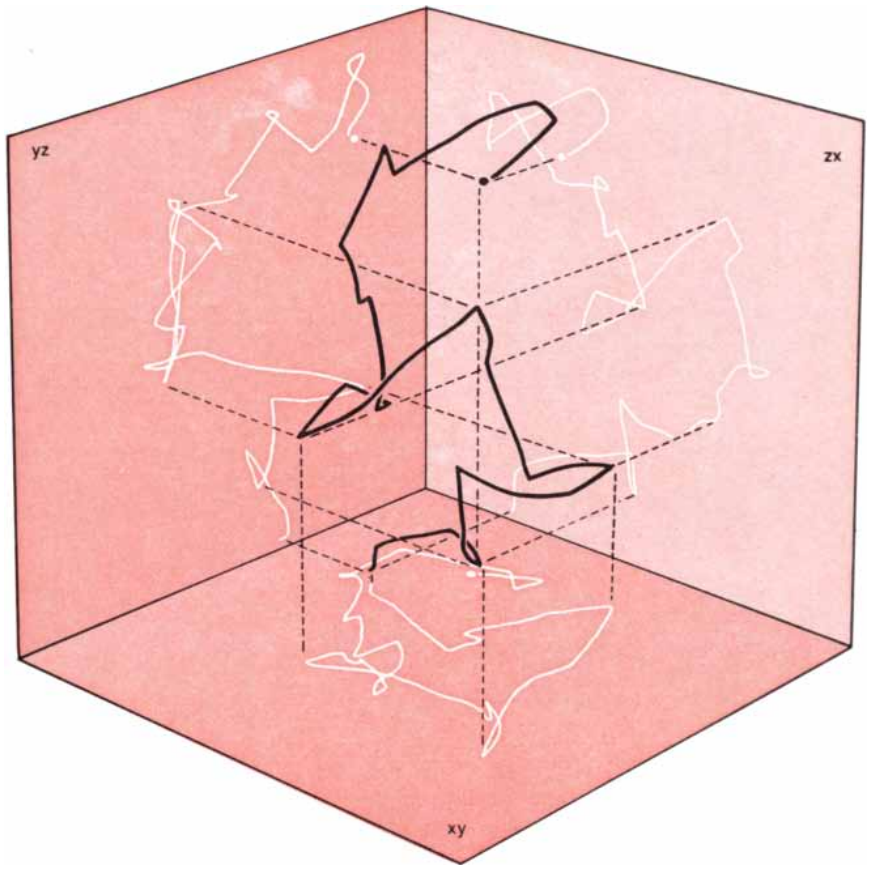
The filaments appear to have no machinery for interconverting chemical and mechanical energy. Bacterial flagella thus differ radically from the much larger and more complexly structured cilia and flagella that propel nucleated cells such as protozoa [see "How Cilia Move," by Peter Satir; SCIENTIFIC AMERICAN, October, 1974]. Yet each segment of the filament must somehow do work as it moves through the viscous environment. This paradox led Aaron Klug of the Medical Research Council Laboratory of Molecular Biology in Cambridge and Jack Lowy and M. Spencer of King's College, London, to propose in 1967 and 1968 that waves of bending are driven from the base, propagated by changes in the relative orientation or the shape of the flagellin subunits. It seemed to me that since passive interactions among flagellin molecules are fundamentally elastic, such waves would have to damp out. My reasoning may be wrong, but



TRACKING MICROSCOPE developed by the author continuously moves a small chamber containing bacteria so as to keep a particular cell fixed in space and its image fixed in the optical eyepiece and on the detector. The instrument produces a record of the cell's movement in three dimensions: three voltages providing measures of the cell's position along the *x*, *y* and *z* axes. The image is split (*top right*) and projected on three pairs of fiber-optics fibers in the detector. Movement of the cell up or down sharpens the image, and thereby increases the amount of light, on one or the other *z*-axis fiber; movement in the horizontal plane varies the light falling on each of the *x* and *y* fibers. The differences between the outputs of the two fibers of each pair are converted into electrical error signals that are amplified to drive the transducer: three sets of coils and magnets similar to those that drive a loudspeaker.

doubts about the feasibility of the wave-propagation mechanism led to work that has helped to establish the validity of the alternative hypothesis: that the filaments rotate rigidly.

Most people have considered the latter possibility to be inherently implausible. It requires the structural equivalents of a rotor, a stator and rotary bearings. Such components seem rather unbiological and are unknown in higher organisms. Bruce A. D. Stocker of the Stanford University School of Medicine did note in 1956 that rotation of a fixed helix was at least conceivable. Raymond N. Doetsch of the University of Maryland argued in favor of the idea in 1966 but then opted instead for a model in which the filaments wobble or pivot. Robert Jarosch of the University of Salzburg took up the rigid-rotation cause in 1967 but then in 1972 described an experiment from which he concluded that an entire flagellar bundle rotated as a unit. William F. Harris of the University of the Witwatersrand has since pointed out a major difficulty. One cannot distinguish rigid rotation from helical-wave propagation by simply looking at a cell; one has to be able to follow the motion of features on the surface of a filament and such features are too small to be resolved in the light microscope [*see upper illustration on page 39*].

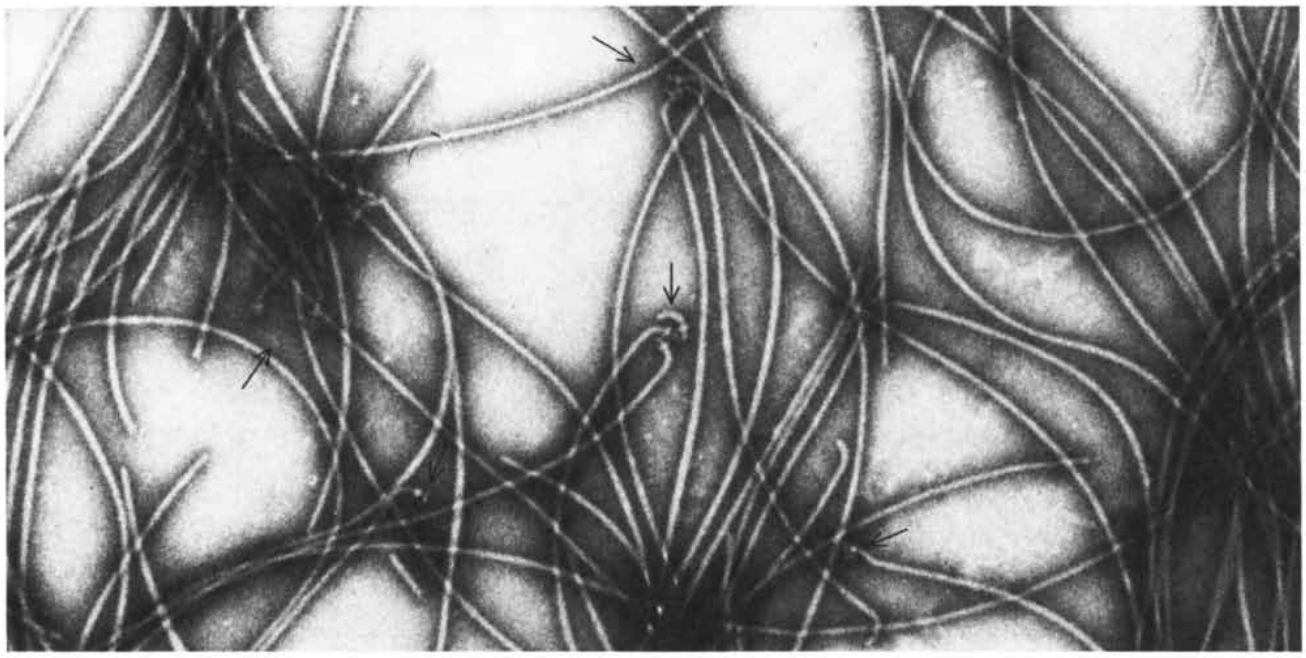


SWIMMING E. COLI that was followed by the tracking microscope covered the path shown here (*black*) in about 30 seconds. The two-dimensional projections of the path on the *xy*, *yz* and *zx* planes are also shown (*white*). While the bacterium was being tracked it executed 26 runs and twiddles; the runs were at a speed of about 20 micrometers per second. The cube that is formed here by the three planar projections is about 130 micrometers on a side.

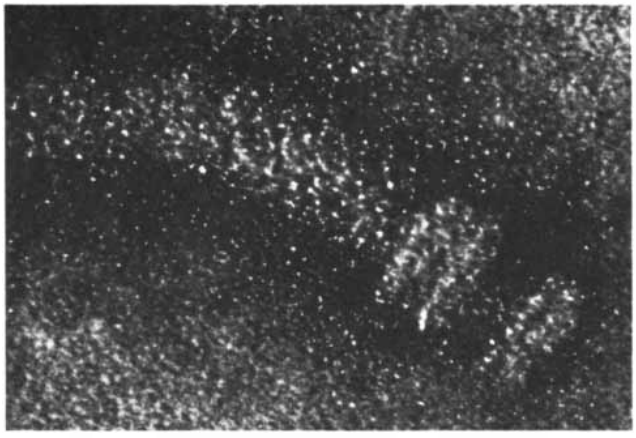
Observable or not, it is clearly impossible for the flagellar bundle of *E. coli* to rotate as a unit because the filaments arise at widely distant sites on the surface of the cell. If the filaments rotate individually within a bundle, however, why do they not become hopelessly entangled? Lowy and Spencer had discussed that point, but I had missed it. It was Robert A. Anderson of the Sandia Laboratories who convinced me that two similar spiral filaments can indeed rotate side by side. He simply bent a pair of wires into spirals and rolled the ends between his fingers. The wires rotated without tangling. (The wires were not exactly helical, but they both had the same shape, and that was the important requirement.) The demonstration suggested a means of distinguishing the rotation of individual filaments from other mechanisms. If the filaments rotate individually, they rotate in relation to one another; if they bend or wobble or rotate as a group, they simply move side by side. The alternatives can be distinguished. Imagine that your wrists are held together by a pair of close-fitting handcuffs or that one wrist is encumbered by a bar projecting from it at a right angle. You could wave or bend your arms, but you could not rotate them (even if you were given rotary joints at the shoulders).

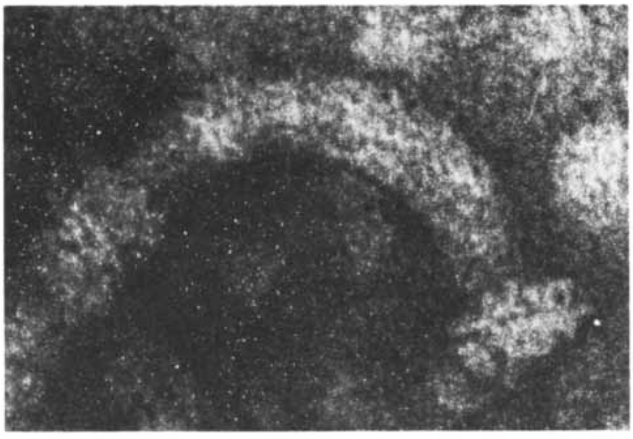
It was known that trace amounts of bivalent antifilament antibodies, which can link adjacent filaments in a bundle, stop bacterial movement (just as handcuffs would stop rotary arm movement). Univalent antibodies, which cannot link adjacent filaments, do not stop the movement even when they are applied in large amounts. Species with a single filament are not affected by either kind of antibody. S. *typhimurium* is stopped abruptly by the bacterial virus χ and the multiply flagellated *Bacillus subtilis* by the bacterial virus PBS1, both of which wrap their tail fibers around a flagellar filament and jut out into the surrounding medium (like the bar on the wrist). A single PBS1 particle can stop a cell even if the viral genetic material has been destroyed enzymatically; the virus must therefore act mechanically rather than by infecting the cell.

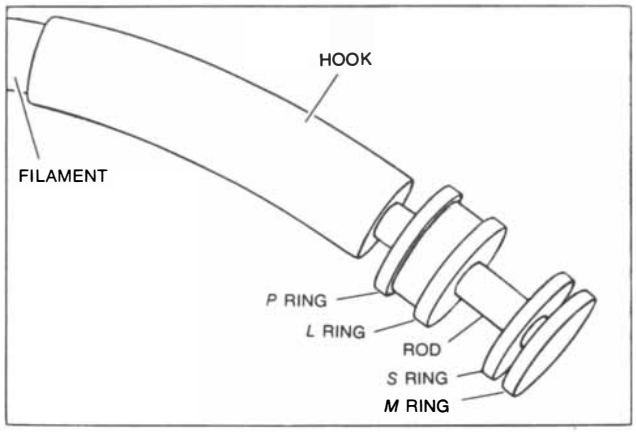
We then learned of an experiment that provided rather direct proof. Michael R. Silverman and Melvin I. Simon of the University of California at San Diego had found mutants of *E. coli* that had abnormally long hooks, which they called polyhooks. The normal hook is a curved segment of the flagellum about .05 micrometer long that joins the proximal end of the filament to the basal structure—a rod and a set of rings—associated with the cell wall and cytoplasmic membrane [*see illustrations on next page*]. The polyhook is a wavy segment between one micrometer and two micrometers long. What appears to be a normal filament is often attached at its far end. Silverman and Simon proved that the polyhooks were made of the same protein as normal hooks by preparing antibodies against normal hooks and showing in electron micrographs that the polyhooks became coated by the anti-hook antibodies. In the course of that demonstration they found that if they added the antibodies to mutants that had polyhooks but no filaments (such mutants do not swim), the cells formed

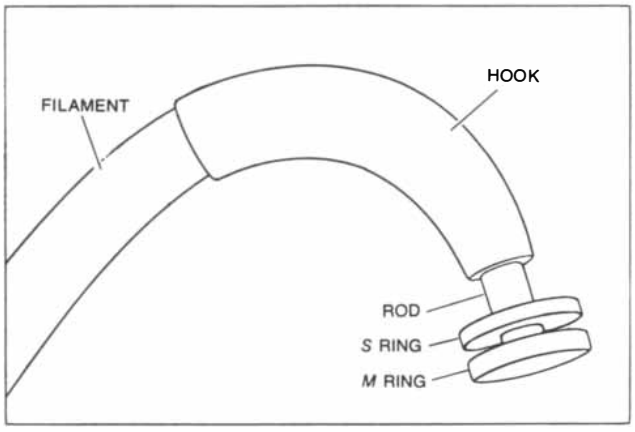


PURIFIED FLAGELLA were isolated from *E. coli* by M. L. DePamphilis, now at the Harvard Medical School, and Julius Adler of the University of Wisconsin. Flagella, negatively stained with uranyl acetate, are enlarged about 70,000 diameters in this electron micrograph, which was made by DePamphilis. A curved "hook" and a basal structure can be seen at one end of many filaments (*arrows*).









HOOKS AND BASAL STRUCTURES of flagella of two kinds of bacteria are enlarged 900,000 diameters in micrographs made by DePamphilis; the elements seen in the micrographs are identified in the drawings. In *E. coli*, a gram-negative bacterium (*left*), four rings are mounted on a rod attached to the proximal end of the hook. The *M* ring binds to a preparation of the inner (cytoplasmic) membrane of *E. coli* and the *L* ring binds to a preparation of the outer (lipopolysaccharide) membrane of the cell wall. In gram-positive bacteria such as *Bacillus subtilis* (*right*) the *L* and *P* rings are not present; the *M* ring binds to the cytoplasmic membrane.

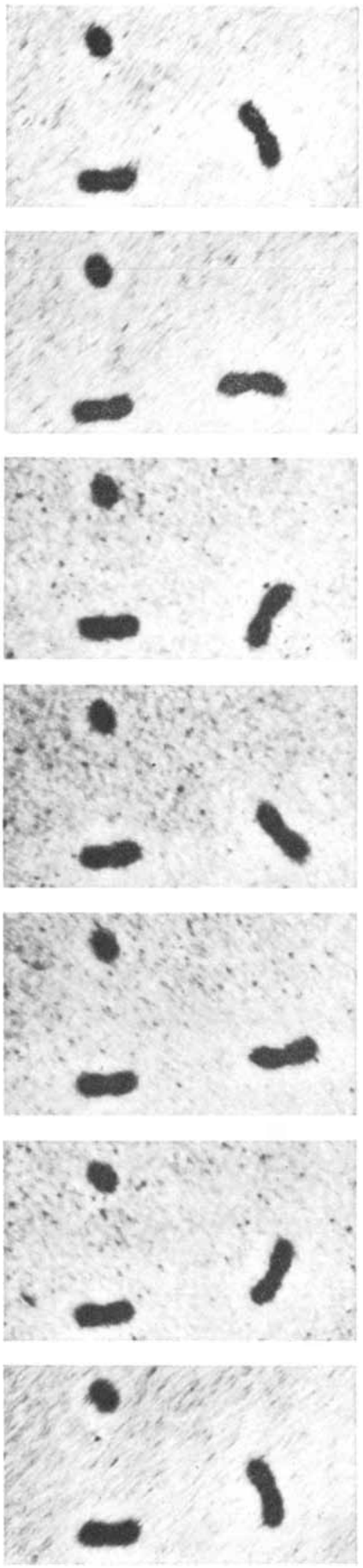
pairs and the two members of each pair rotated in opposite directions. Such counterrotation is precisely what one would expect if the hooks are driven by rotary motors at their base and if the antihook antibodies link a single polyhook from one cell to a polyhook (or several polyhooks) from another cell.

We concluded that the evidence favored a model in which each filament rotates rigidly. We noted that a rotary motor could be built from the elements known to exist at the base of the flagellum, and we suggested that the bending of the filaments around the sides of the cell (required for the formation of a bundle) is facilitated by the hook, which serves as a flexible coupling or universal joint.

Silverman and Simon have gone on to show that the body of a polyhook mutant or of a mutant that has straight filaments will spin if it is tethered to a glass slide with antihook or antifilament antibodies. Both kinds of mutant are nonmotile when they are free, but when the hook or filament is linked to the glass, they rotate—now clockwise, now counterclockwise. The cell body does not merely precess or wobble; it clearly rotates [*see illustration at right*]. (One could in principle repeat the experiment with an electric fan; if the blades of the fan were fixed to a wall and the motor were not encumbered by the base or the power cord, the motor would spin around.) The tethering of an otherwise nonmotile cell makes evident the rotation that, as Harris noted, cannot ordinarily be observed unequivocally. Silverman and Simon found still another way to visualize the rotation. They link polystyrene-latex beads to polyhooks or filaments by means of antibodies. When this is done to a cell with straight filaments and the cell is examined in the light microscope, often several beads can be seen in tangential contact with a line (an invisible one) projecting from the cell surface. The beads remain fixed in relation to one another as they revolve together around the line, and the body of the cell rotates around the same axis in the opposite direction [*see upper illustration on next page*].

Steven H. Larsen and others in Adler's laboratory have extended the tethering technique to show that the twiddle in *E. coli* is initiated by a change in the direction of rotation of the flagella. Mutants that run incessantly (that do not twiddle) rotate in one direction when they are tethered; mutants that twiddle incessantly (that do not run) rotate in the opposite direction. When wild-type bacteria are tethered and stimulated in a manner that would cause them to run (if they were free), they rotate for a time in the same direction as the incessantly running mutants; when they are stimulated in a manner that causes them to twiddle (when they are free), they rotate briefly in the opposite direction. The shafts of the flagellar motors turn counterclockwise (viewed from outside the cell) when the cells run and clockwise when they twiddle. Since the normal flagellar filaments of *E. coli* are left-handed (an object moving away from an observer along such a helix goes counterclockwise), these results are consistent with the observation, discussed above, that the cell runs when it is pushed by its flagellar bundle. The twiddle in *E. coli* and the shock reaction in *C. okenii* appear to be analogous events: both begin and end when the filaments change their direction of rotation. As a matter of fact, one can show with wire models, as Lowy and Spencer did, that the rotation is less successful when the filaments rotate so as to pull the cell; the wires tend to tangle. This may explain both the spontaneous termination of the shock reaction and the brevity of the normal twiddle.

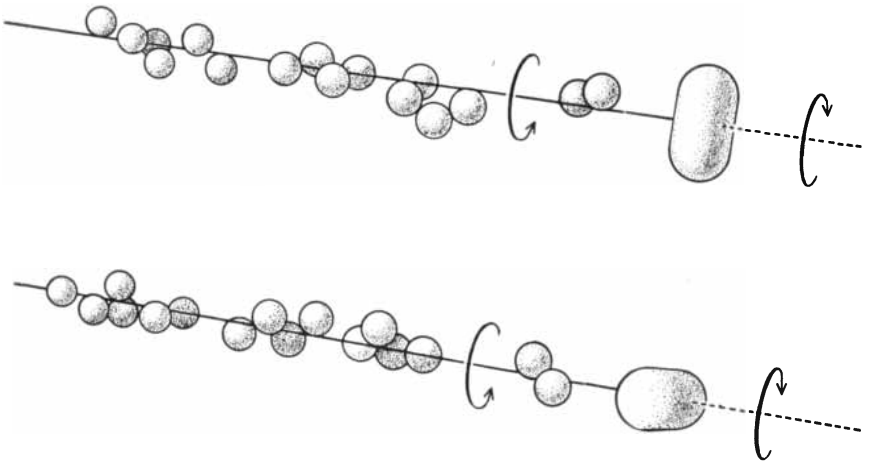
Larsen and his colleagues have also demonstrated that adenosine triphosphate (ATP) is not the immediate source of energy for flagellar rotation. This is surprising in view of the involvement of ATP in other motile activities, including the contraction of muscle and the bending of cilia. Mutants of *E. coli* that are unable to use the high-energy intermediate in oxidative phosphorylation for synthesizing ATP, or to use ATP for generating the high-energy intermediate, fail to swim when they are grown without oxygen, even though they are producing ATP. They do swim in the presence of oxygen, even if their ATP has been destroyed by treatment with arse-



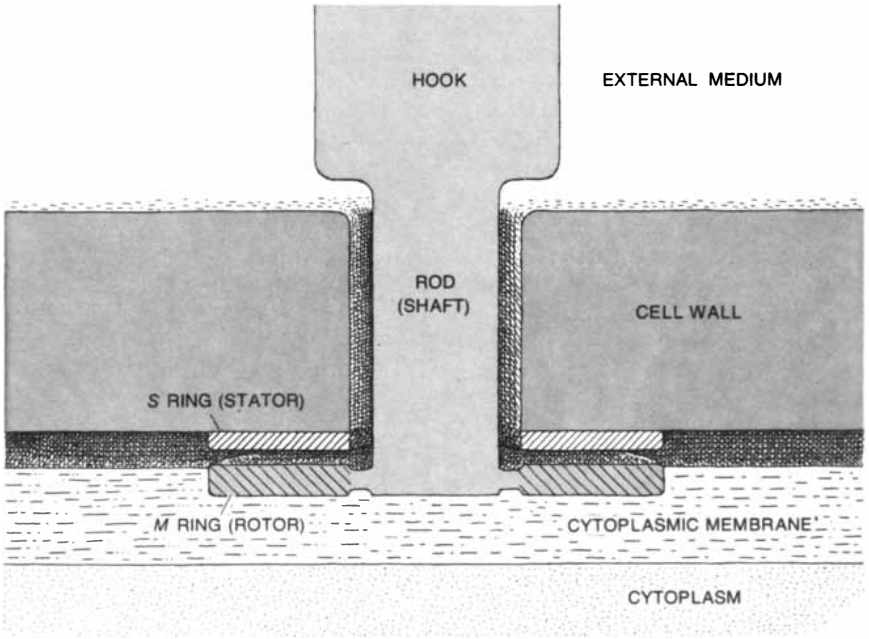
TETHERED E. COLI makes one complete rotation in seven frames of a film made in the author's laboratory. The bodies of the cells at the left are stuck to the cover slip of the microscope slide. One filament of the cell at the right is fixed to the cover slip by antifilament antibodies; the cell body rotates counterclockwise. Its changing orientation is particularly evident because it is asymmetrical, having been washed free of growth medium when it was about to divide. The flagellum responsible for the motion is in the lower part of the cell (*in top frame*).

nate. The energy for motility must therefore be derived from the high-energy intermediate rather than from ATP.

It is possible to monitor the rotation of a tethered cell with the tracking microscope, which follows the point on the edge of the cell body closest to the center of the tracking chamber. If the cell spins clockwise, the transducer traces out a clockwise path; recordings of the transducer's velocities along the x and y coordinates look very much like sine waves. The two traces are 90 degrees out of phase, and the sign of the phase difference indicates the direction of rotation. Reversals are seen to come quite abruptly. The cell spins in either direction at about the same speed. Its motion is remarkably smooth. I have not been able to find any evidence for rotational Brownian motion, which would indicate passive slip between the filament and the cell wall (as in a fluid drive). Nor have I found evidence for discrete steps, which would indicate a ratchetlike drive; if there are steps, they may simply be so small that they are not reflected in the traces.



ROTATION CAN ALSO BE VISUALIZED by coating microscopic beads with antifilament antibodies so that the beads become linked to a filament of a mutant of *E. coli* that has straight filaments rather than helical ones. The filament is not visible in the light microscope; the beads seem merely to be in tangential contact with a line projecting from the cell. They revolve around that line in one direction while the cell body rotates in the other direction, as is shown in these two successive views. Now and again the beads slow down, stop and then start up in the opposite direction; the cell body reverses its rotation synchronously.



ROTARY MOTOR, a model proposed by the author, is shown as it might appear in a gram-positive bacterium. The flexible hook serves as a universal joint coupling the rod to the filament. The torque is generated between the *M* ring, which is mounted rigidly on the rod and rotates freely in the cytoplasmic membrane, and the *S* ring, which is mounted rigidly on the wall. Torque could be generated by the translocation, through the cytoplasmic membrane and the *M* ring, of ions that interact with charges on the surface of the *S* ring. The additional pair of rings in a gram-negative bacterium may serve as a bushing for the rod's passage through the cell's more complex wall. The rings are about .02 micrometer in diameter.

The evidence at hand suggests a model for the rotary motor in which the torque is generated between two elements in the basal body, the *M* ring and the *S* ring [*see lower illustration at left*]. The rod (which is connected to the filament by the hook) is fixed rigidly to the *M* ring, which rotates freely in the cytoplasmic membrane. The *S* ring is mounted on the cell wall. (Note that the motor must be mounted rigidly somewhere on the cell wall if the torque is to be applied.) The torque could be generated by the active translocation of ions through the *M* ring to interact with charged groups on the surface of the *S* ring. A number of membrane transport processes are known that do not involve ATP. The direction of the rotation would be determined by the timing of the ion flow. This would require some kind of molecular control, a control that could be modulated in turn by signals generated by the chemoreceptors.

The implausible, then, appears to be true. Bacteria swim by rotating their flagella. They alter course by changing the direction of the rotation: cells with polar flagella back up; cells with lateral flagella try a new direction at random. The probability of the occurrence of these events is biased by sensory reception: the cells tend to move toward regions they find more favorable. The flagellar motor and the sensory machinery may soon be understood in molecular detail.

Having glimpsed and described the cilia and flagella that propel protozoa, Leeuwenhoek—ever amazed and even moved by the small size of complex things—went on to report: "I see, besides these, other living animalcules which are yet more than a hundred times less, and on which I can make out no paws, though from their structure and the motion of their body I am persuaded that they too are furnished with paws withal." The "paws" are beginning to yield up their secrets.

For those who want to do everything right...

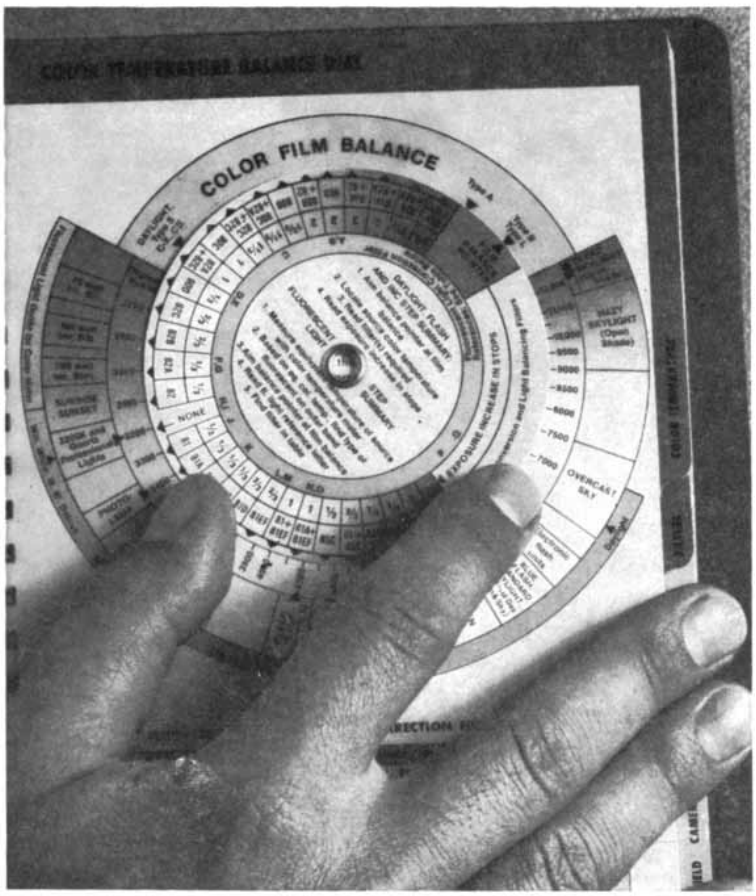
exactly right

"Do" is the operative word.

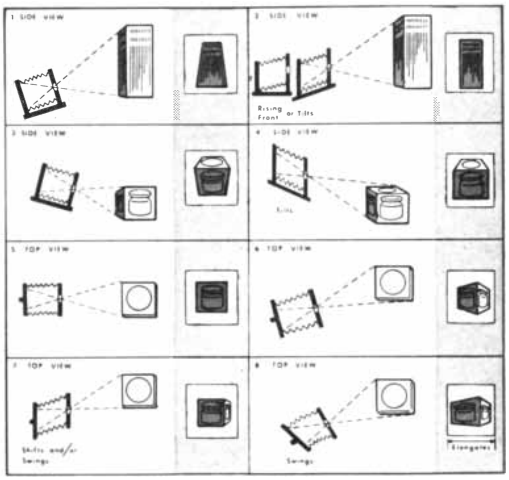
You are constantly being promised, in effect, "Our product gives excellent results with very little effort." As much as anybody offering any kind of merchandise, we have been believed. We would not have fared as well as we have if so many people the world around had not felt satisfied.

Some there are, however, to whom such a promise borders on affront. They take no joy in avoiding effort. They think that more effort is apt to bring forth more excellence. In general, they are right. Them we also serve.

The new Kodak publication No. R-28 is not just words and illustrations. Some of its pages are equipped with dials such as this one, for a photographer who cares very deeply about color temperature of illuminants.



Below is a sample of how another page guides control of perspective and depth of field by means of the adjustments for which the view camera—horrible antique!—retains its value.



Still another page carries gray scales and color patches of some considerable precision in light reflectance.

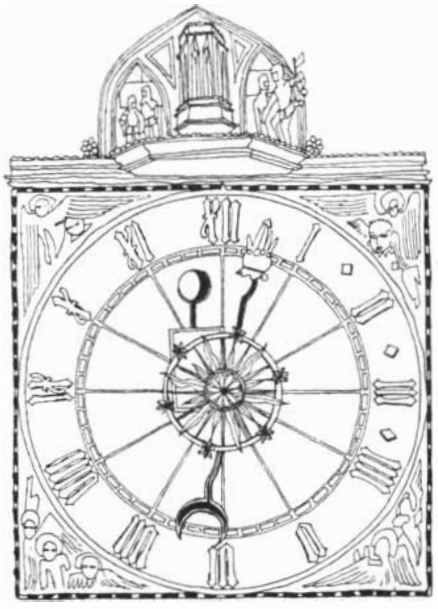
Other topics on which the reader is spared few details of working significance: film selection and care, exposure determination, filters, flash, camera lenses and shutters.

Actually, nobody reads a work like this. One merely uses it. Supplements one's instincts and training. Particularly if the training has had to cover other subjects than photography but photographic ambition runs high.

The companion Kodak publication No. R-19, with such things as an apertured page containing viewing filters and multi-dial computers riveted into the pages, deals with color photography *after* the shutter has clicked.

Order these books where you get your photographic supplies or from a bookseller.

Kodak



SCIENCE AND THE CITIZEN

Lost Opportunity

The long awaited international conference held in Geneva in May to review the operation of the Treaty on the Nonproliferation of Nuclear Weapons appears to have accomplished little or nothing toward preventing the imminent spread of nuclear-weapons technology around the world. Indeed, the fundamental disagreement that had arisen earlier between the nuclear "have" and "have not" nations over the proliferation issue seems if anything to have been exacerbated by the disappointing outcome of the conference and by the trend of subsequent events. Only 57 of the 96 current parties to the treaty participated in the month-long meeting, and no single group of nations could muster the two-thirds majority needed to recommend measures aimed at strengthening or expanding the provisions of the treaty. The conference ended with the adoption of a compromise declaration that essentially repeated the pledges of the original treaty and urged the parties to try harder; it also committed the parties to hold another such review conference in 1980. Even this "consensus" statement was hedged with official reservations by many of the parties.

Critics of the position taken by the two major nuclear powers and their allies before, during and after the conference have argued for some time that the main shortcoming of the treaty has been the failure of the nuclear-weapons states to live up to specific obligations undertaken by them at the time the treaty was negotiated. They cite, for example, Article IV of the treaty, which obligates the parties "to facilitate ... the fullest possible exchange of equipment, materials and scientific and technological information for the peaceful uses of nuclear energy ... especially in the territories of the non-nuclear-weapon States Party to Treaty," Article V, which provides that under appropriate international supervision "potential benefits from any peaceful applications of nuclear explosions will be made available to non-nuclear-weapon States Party to the Treaty on a non-discriminatory basis," and Article VI, which pledges the nuclear-weapons parties "to pursue negotiations in good faith on effective measures relating to cessation of the nuclear arms race at an early date and to nuclear disarmament." These three provisions, and particularly Article VI, it is maintained, constitute the *quid pro quo* that induced the non-nuclear parties to agree not to acquire nuclear weapons in the first place. In effect, then, the Nonproliferation Treaty is seen by many as a contract in which the non-nuclear parties undertook to halt the "horizontal" proliferation of nuclear weapons (that is, their spread to additional countries) in return for the pledge of the nuclear parties to halt the "vertical" proliferation of nuclear weapons (that is, their further development and deployment).

In spite of this widely held view of the contractual nature of the Nonproliferation Treaty and the corollary conviction that the major nuclear powers had failed to uphold their end of the bargain, the nuclear-weapons states and their allies rejected all the major proposals made at the review conference by the non-nuclear countries, led chiefly by Mexico. The rejected proposals included measures calling for an end to underground nuclear tests; for a reduction in nuclear arsenals; for a pledge on the part of the nuclear powers not to use or threaten to use nuclear weapons against non-nuclear parties to the treaty; for substantial aid to the developing countries in the peaceful uses of nuclear energy; for immediate negotiations leading to the creation of a special international regime to carry out peaceful nuclear explosions; for equal treatment in the application of nuclear safeguards, and for an agreement to respect all nuclear-free zones.

According to William Epstein, a special United Nations consultant on disarmament who attended the review conference in an unofficial capacity, "it is almost incredible that the nuclear powers, whose interest in the survival of the Nonproliferation Treaty is far greater than that of the non-nuclear powers, and who are the delinquent parties, could have been so insensitive to the legitimate demands of the non-nuclear powers." The two superpowers and their "somewhat embarrassed" allies, he reports, "were interested only in technical questions: tightening international controls and safeguards on exports and on the peaceful activities of the non-nuclear-weapon states, while refusing to accept them fully for themselves, something the non-nuclear parties regarded as a form of nuclear neo-colonialism."

A Plan for Economic Growth

Concern about the instability of the U.S. economy and the failure of short-term remedies to deal effectively with inflation and unemployment has generated a movement in Congress to require rational economic planning at the national level. A "Balanced Growth and Economic Planning Act" was introduced in May by senators Hubert H. Humphrey and Jacob K. Javits and has since attracted a dozen more sponsors. The Congressional Joint Economic Committee held two days of hearings on the bill in mid-June.

The bill would establish an Economic Planning Board whose primary function would be to prepare every two years a "balanced economic growth plan," a blueprint for Federal Government actions affecting the economy. The plan would establish economic objectives, with particular attention to attaining "full employment, price stability, balanced economic growth, an equitable distribution of income" and certain other goals. It would identify the requirements for achieving such goals and recommend necessary legislative and administrative actions with regard to money supply, the Federal budget, interest rates, taxation and so on. After approval by an interdepartmental Council on Economic Planning the plan would be submitted to Congress and at the same time transmitted to state and local officials for comment. After hearings by the Joint Economic Committee, Congress would pass a concurrent resolution either approving

the plan, with or without modification, or disapproving it; if disapproved, the plan would be modified and resubmitted. As finally accepted and promulgated the plan would become the basis of economic policy: the President would be expected to "ensure" that Federal agencies further the plan's objectives and to "encourage" state and local governments and the private sector to do so.

At the June hearings proponents of the bill pointed out that individuals, families, small and large businesses and institutions all plan their economic activities; the Federal Government, which has the most power to influence the economy, is the last bastion of unplanned activity. The complex U.S. economy cannot be dealt with haphazardly, they said; piecemeal legislative or administrative action designed to solve one problem often creates another. The establishment of a planning board would not give the Government any new powers to regulate or direct the economy, proponents argued; the plan would be responsive to public opinion and subject to Congressional control. The bill, it was said, not only would rationalize the Government's economic activities but also would make economic policy explicit and expose it to much broader public debate and control than at present.

Opponents of the bill insisted that the normal give-and-take of the marketplace remains the best regulator of economic activity and that a national plan would infringe on personal liberty and be the first step toward the destruction of private enterprise. Defenders of the bill reply that lack of planning does not prevent Government regulation; it merely makes that regulation unjust or ineffective. There is no reason to believe, they say, that basing Government policy on rational planning will make it more intrusive.

The Missing Neutrinos

It is generally believed that the source of the sun's energy is the proton-proton chain, in which four hydrogen nuclei (protons) fuse in a series of steps to form a nucleus of helium. In some of the steps the nuclei of heavier elements, principally the isotopes beryllium 7 and boron 8, are briefly created. These nuclei decay, emitting, among other things, energetic neutrinos, some of which should be detectable on the earth. In 1968 Raymond Davis, Jr., of the Brookhaven National Laboratory set up an ambitious experiment to try to detect some of the solar neutrinos. His results over the past three years have shown that the number of neutrinos reaching the earth may be as little as a thirtieth of the number predicted by theory. They have therefore raised the question of whether there is something fundamentally wrong with the proton-proton theory. In a recent paper in *The Astrophysical Journal* Fred Hoyle has proposed a possible explanation of why the sun is emitting so few neutrinos that still assumes that the proton-proton mechanism is the right one.

Hoyle suggests a new model of the sun that differs from the standard model in two important ways: first, the core of the sun represents between 30 and 50 percent of the sun's total mass (about two-thirds the amount assumed in the standard model), and second, the core is far richer in heavy elements and in hydrogen (in the standard model the core is mostly helium, with only about 10 percent hydrogen).

The external layers of the sun are observed to be 70 percent hydrogen, 28.5 percent helium and 1.5 percent elements heavier than helium. The core in Hoyle's model is the result of a mixture of two types of material; one type was originally 96 percent hydrogen and 4 percent helium and the other was 50 percent hydrogen and 50 percent elements heavier than magnesium. Hoyle's choice of compositions is based on recent work of his suggesting that the material out of which stars form is far from uniform in composition. He proposes that material of these particular compositions is what gave rise to the sun's core.

At first the core did not manufacture much helium by the proton-proton chain. It was also not very luminous because its gas, rich in heavy elements, was quite opaque to its radiation. It was only after the core had already been formed that it swept up the material in space that became the sun's exterior layers. For the core to have enough internal pressure to support the new exterior layers the temperature at its "surface" had to be about 10 million degrees Kelvin. At that temperature, however, the amount of energy radiated by the exterior layers was much greater than the amount radiated by the core because the exterior layers, poor in the heavy elements, were much more transparent to the radiation. Therefore to ensure that the flow of energy from the core to the exterior was continuous the nuclear reactions within the core had to proceed at an increased rate. Moreover, for that energy to be transported through the opaque core to the exterior layers, convection currents had to develop within the core.

What is the effect of Hoyle's model on the production of neutrinos in the solar interior? First, the model requires that the core contain between 70 and 75 percent hydrogen. The amount of boron 8 manufactured in the proton-proton chain is sensitive to the composition of the core, and it is very small in the presence of so much hydrogen. Since the decay of boron 8 is one of the principal sources of energetic solar neutrinos, if there is very little boron 8 produced in the sun, then there will be very few neutrinos emitted. Second, convection currents within the core distribute its constituents uniformly. The decay of beryllium 7, the other reaction by-product of the proton-proton chain that is an important source of neutrinos, is strongly dependent on density. Therefore the rate of decay decreases when the beryllium is transported to the edges of the core. Between these two effects the number of neutrinos produced in the core of Hoyle's model of the sun very closely matches the number that has been calculated from the neutrinos observed to reach the earth.

Computerized Elections

The advance of the computer in tallying the results of elections has reached a stage where nearly 15 percent of the voters in the U.S. now receive ballots that can be read by machine. The Clearinghouse on Election Administration, a Federal agency, is concerned about making sure that computerized elections run smoothly; it is also uneasy about the possibility that computerization will diminish the control of election officials over voting procedures, thereby increasing the likelihood of vote fraud. The Clearinghouse has therefore asked the National Bureau of Standards for advice on dealing with these concerns. A summary of a year's study by the Bureau's Institute for Computer Sciences and Technology appears in *Dimensions*, the technical news bulletin of the Bureau of Standards.

In a typical election involving a computer the voter receives a ballot that is actually a punch card. He votes by punching out a precut slot associated with his candidate or his view of an issue. His card, along with the other cards, is read by a punch-card reader at a counting center, and the information is fed into a computer that is programmed to count allowable votes and to print out the results. (The ballot can also be a card that is marked with a dark pencil or with a special ink; in any case the marking can be read by machine.)

In studying a number of computerized voting systems the Bureau of Stan-

dards found no evidence of fraud, but it did find instances of management failure (causing things to go wrong in computerized voting) and of insufficient attention to security, so that the opportunity for fraud was enhanced. The Bureau recommended that computerized voting programs be thoroughly tested in advance. It also recommended schemes of checking to ensure that all ballots are accounted for and that the computer is not erring in its count. Moreover, the Bureau said, to maintain the security of vote-counting programs a computer employed to count votes should not be used for anything else.

Computerized Cars

In response to the energy shortage, antipollution regulations, increasing public concern about vehicle safety and, not least, decreasing sales, the automobile industry seems ready to make greater changes in its cars in the next five years than were made in the previous 25. The key element in these innovations will be the electronic microprocessor, the central processing unit developed for microcomputers, contained on a single "chip" of silicon about a tenth the size of a postage stamp (see "Microcomputers," by André G. Vacroux; SCIENTIFIC AMERICAN, May). The role of the microprocessor will be to collate the inputs from dozens of sensing elements monitoring engine conditions, exhaust emissions, the general status and performance of the vehicle, road and traffic conditions and even the state of the driver's alertness. The microprocessor will initiate changes as necessary to achieve optimum vehicle performance and will actuate a variety of digital displays, including such things as a road-surface indicator, miles remaining on a trip, estimated time of arrival and fuel consumption both for a trip and for any given moment. In a recent issue of the *IEEE Spectrum* Ronald K. Jurgen lists more than 60 tasks that might be handled by automotive microprocessors. Some of the more exotic would be automatic "station-keeping" in highway traffic (maintaining a safe distance from the car ahead), the prediction of an imminent crash and the continuous monitoring of the driver's brain waves to detect impending drowsiness or other undesirable mental states.

According to Jurgen, General Motors has been studying automotive electronic systems under four headings coded Alpha, Sigma, Delta and Beta. The Alpha series is concerned with overall system integration, the Sigma series with display systems, the Delta series with diagnostic functions and the Beta series with driver physiology. The findings will ultimately be integrated into an Omega system. In Alpha IV, the fourth version of the Alpha series, system integration was handled by a single-chip microprocessor with a four-bit word length. Its functions included the control of ignition timing, ignition "dwell" and traction level and the prevention of wheel lock, along with digital displays such as the trip odometer, an elapsed-time indicator and an engine-speed indicator.

Ford is emphasizing a digital control system that maximizes fuel economy and drivability for a given permissible level of exhaust emission. The microprocessor in the Ford system requires a program of 1,500 12-bit words stored in a read/write semiconductor memory.

Chrysler, according to Jurgen, will introduce a computer-controlled engine in some of its 1976 models that will improve fuel economy and eliminate the need for a catalytic converter to clean up the exhaust gases. The computer will make it possible for the engine to run on air-fuel ratios of 18 : 1 or more instead of the conventional 15 : 1 or 16 : 1. The leaner mixtures should improve fuel economy some 5 percent.

Meanwhile engineers at the Massachusetts Institute of Technology have recently developed systems that will constantly monitor engine speed, power output and power demand and in thousandths of a second make optimum adjustments in engine functions such as fuel-feed rate, fuel-air mixture, spark timing and perhaps even gear shifting. The heart of the system, developed at M.I.T.'s Innovation Center, is a torque meter that can measure variations in crankshaft torque with a sensitivity of one foot-pound in 1,000. The torque is measured by an optical system that can determine a twist in the shaft of as little as 15 millionths of an inch over a distance of nine inches. The electronics in the torque sensor are said to cost about $10. The complete engine-control system, including the microprocessor, should cost less than $50 in large-scale production.

Drop by Drop

Irrigation plays an essential role in world agriculture, but it is a notably inefficient procedure. Most of the water that is conveyed to the plants through ditches and sprinklers is lost by evaporation before it can reach the plant roots. A new method of irrigation developed in Israel that requires substantially less water to produce the same yield as conventional irrigation is rapidly being adopted in many parts of the world. The method delivers the water to the plant one drop at a time. The water infiltrates the soil at the base of the plant and moves directly to the root zone. Losses by evaporation are kept to a minimum and there is no runoff.

Drip irrigation grew from a chance observation made in the mid-1930's. An Israeli hydraulic engineer, Simcha Blass, noticed as he was passing a tree-lined fence that one tree was much taller than the others. He became curious and took a closer look. He discovered that although all the trees were being irrigated, the taller one was also being watered by a constant drip from a leaking water-pipe connection.

A modern drip-irrigation system consists of a plastic-pipe distribution network with drip emitters at the base of each plant. Although the installation cost is high, the system can be made to operate automatically, thus decreasing the labor cost of irrigation. Under some circumstances fertilizer and pesticides can also be applied through the drip system. A major operating problem is clogging of the drip emitter.

Although drip irrigation is designed primarily to save water, it has the additional advantage of increasing plant vigor and improving crop quality. The method has been successfully applied to orchard trees and to greenhouse vegetables. Its greatest economic benefit, however, may be in its future application to row crops such as cotton and maize.

Research at the Texas Water Resources Institute of Texas A&M University has shown that drip irrigation yields water savings of 50 percent or more compared with conventional irrigation. Another conservation asset is that in sandy soils somewhat brackish water can be used. The salt in the water is concentrated at the outer edge of the wetted zone, whereas in conventional irrigation the salt is concentrated in the surface layers and carried down to the root zone with subsequent applications of water.

Origins

At age 16 Albert Einstein discovered a paradox that, as he later wrote, contained "the germ of the special theory of relativity." He stated the paradox in this way: "If I pursue a beam of light with the speed c (speed of light in a vacuum), I should observe such a beam of light as a spatially oscillatory electromagnetic field at rest. However, there seems to be no such thing, whether on the basis of

experience or according to Maxwell's equations." Today the paradox is usually avoided by pointing out that it is impossible to travel at the speed of light. Banesh Hoffmann of Queens College of the City University of New York has recently suggested that the paradox is much subtler.

The differential equations devised by James Clerk Maxwell describe a light wave in terms of the vectors of its electric and magnetic fields. The wave vectors can be transformed by the methods of classical, nonrelativistic physics to describe the wave as it is seen by an observer moving with speed c, and the result is the arrested oscillation Einstein envisioned. The transformed vectors, however, no longer satisfy the Maxwell equations, which seems to confirm Einstein's insight.

Hoffmann shows that the wave vectors fail to solve the Maxwell equations not only when the observer moves with a speed of c but also when he moves with any speed at all. An observer strolling at a few centimeters per second confronts the same contradictions as one moving at the speed of light. Thus it would seem that the observer cannot move at all! This apparent absurdity is a result of applying a classical transformation to the wave vectors but a relativistic one to the Maxwell equations. It derives from the fact that Maxwell's equations were conceived as pertaining to a luminiferous ether defining a frame of reference that can be considered to be at rest. Given these assumptions, the paradox can be avoided. It is necessary only to transform the equations classically, in the same way that the vectors have been transformed. This procedure introduces into the equations terms that explicitly refer to the motion of the observer, so that the principle of relativity no longer holds.

Writing in *Transactions of the New York Academy of Sciences,* Hoffmann argues that the young Einstein did not pursue the paradox to this length. He does not seem to have seriously considered the case of an observer moving with a velocity between zero and c. If he had, he would have found that mathematically there was no contradiction: the observer would be free to pursue the light beam. Had Einstein considered the case of an observer moving with an intermediate velocity, his problem would have lost its sense of paradox. Intuition apparently told him that the paradox was important, and Hoffmann suggests that the threat of its loss created in the young Einstein a psychological block that prevented him from analyzing his argument too closely.

THE CAUSES OF BIOLOGICAL DIVERSITY

Within a single species individuals differ in many genetic traits, such as the chemical composition of their proteins. It appears that this diversity is actively maintained by natural selection

by Bryan Clarke

In 1636 Sir Thomas Browne wrote: "It is the common wonder of all men, how among so many millions of faces, there should be none alike." Individuality is still a wonder, and the questions it provokes are relevant to matters far removed from human physiognomy: sheep are distinguishable to the shepherd, as fruit flies are to the geneticist. The basis of individuality is variation. If men or sheep or fruit flies can be recognized as individuals, then it follows that their populations exhibit variation. Variation is the material of science, and variation among the members of a species is the material of genetics.

A classical treatment of variation within (and between) species is Charles Darwin's theory of evolution by natural selection. Variation is essential to the theory—it is what natural selection acts on—but for many interpreters of Darwin it is a rare and short-lived phenomenon. Indeed, it has often been suggested that at any given time the individuals making up a population should be almost identical in genetic composition. One of the principal challenges to genetics and ecology during the past decade has been the apparent contradiction between that view and observations suggesting a high incidence of genetic variation in natural populations.

Darwin's theory has great stature in biology, and it has had a powerful influence on other fields of thought. Today it is often regarded as established doctrine. Nevertheless, we have recently come to believe that one aspect of the classical Darwinian canon is incorrect. The error (if it is an error) is more properly attributed to some of Darwin's interpreters than to Darwin himself, but it can be corrected only by revising ideas that have come to be associated with his name.

Classical Darwinians have argued, with reason, that at any one time the majority of species are well adjusted to the demands of their environment. The adjustment, or adaptation, is a result of natural selection. At irregular and infrequent intervals individuals are born that manifest inherited modifications of the general type, or mutations. Because the species is finely adapted to its environment the majority of mutations are disadvantageous and are more or less rapidly eliminated by natural selection. Only rarely is a mutation advantageous, but when one is, it spreads through the population and replaces the previously dominant type.

Mechanism of Mutation

Through the work of Gregor Mendel (completed in Darwin's lifetime but virtually unknown until the 20th century) and the more recent discoveries of molecular biology, we can describe the mechanism of mutation in great detail. We know that the genetic complement of an individual can be described in terms of discrete units (genes). Each gene controls a specific set of developmental processes, and each occupies a particular locus, or place on a chromosome; when there are several forms of a single gene in a population, each form is called an allele.

A gene is now known to represent a segment of DNA, and the genetic locus to represent the position of the segment on a much larger DNA molecule. With few exceptions the product of a gene is a protein, and the development of the organism depends on the particular proteins specified by its genes. A mutation in a gene can produce a change in the corresponding protein; if the mutation occurs in the reproductive cells of the organism, it will be passed on to the next generation.

The traditional Darwinian argument suggests that at any given genetic locus most members of a population carry the same allele: the one that best fits them to their environmental circumstances. The exceptions are the few individuals that carry other alleles at the locus as a result of mutation. In almost all cases the mutant forms have a reduced probability of survival or a reduced reproductive capability. (There are many unfavorable alleles in the human population, including those responsible for such diseases as phenylketonuria and Huntington's chorea. Fortunately each of these alleles individually is rare. It is because the great majority of mutations produce some disability that we are concerned about agents such as ionizing radiation that increase the mutation rate.)

This outline of the received version of the Darwinian theory is perhaps a simplistic one. It should be clear, however, that classical Darwinism, as it has been interpreted, requires that there be rather little genetic variation within a natural population. In fact, however, a great deal of variation is commonly observed.

One way to reconcile theory and observation is to assume that a large proportion of the variation in man and other species is caused by environmental factors rather than genetic ones. At birth—or at conception—we are for the most part alike, it could be argued, and we acquire our diversity only later, through exposure to diverse circumstances. Our faces are the product of our fortunes.

This assertion can be tested by estimating how much of the observed variability is related to parentage rather than to environment, and thereby determining what proportion of the variability is genetic. The method of measurement depends on estimates of the extent to which relatives resemble one

SHELL MARKINGS of the land snail *Cepaea nemoralis* display great variation even though all the snails are members of the same species. Among the most prominent variable traits are the color of the shell and the pattern of dark stripes or bands. Both characteristics are genetically determined and each is inherited independently, so that a shell of any given color can appear with many different banding patterns. Populations in which two or more such varieties are relatively common are said to manifest genetic polymorphism.

GENETIC POLYMORPHISM in another species of land snail, *Partula suturalis*, is expressed by variations in shell coloration and by differences in the direction (left or right) in which the helical shell is twisted. The two variable traits are both inherited, although the direction in which the shell coils is determined by genes carried by the snail's female parent rather than by those of the snail itself.

COLOR AND BANDING of *C. nemoralis* differ from snail to snail within a population. Among the common colors are yellow (*top and middle*) and pink (*bottom*); a shell of either color can have as many as five bands (*top*) or it can be lacking in some or all (*middle and bottom*). Photographs were made by David T. Horsley of the University of Nottingham.

another. For a given character, such as human height or body weight, we can calculate the degree of correlation between parents and their offspring, between siblings and between more distant relatives such as cousins and grandchildren. Knowing the extent of variation for the character in the population—the range of values encountered—we can estimate the proportion of the variation caused by inherited factors, the proportion caused by environmental factors and the proportion that can be attributed to interaction of the two. The result is a measure called heritability, which is defined as the proportion of the variation that can be attributed to the action of genes behaving in an additive manner. Since some genes may interact with one another and with the environment in complex, nonadditive ways, heritability measurements tend to underestimate the genetic component of diversity.

In spite of this tendency toward underestimation, the calculation of heritabilities for some human traits yields quite high values. Human height, for example, seems to have a heritability of from 50 to 80 percent. (It should be emphasized that these values apply only within the Western societies in which they were measured. The heritability of a trait may vary from one population to another, and it certainly depends on the degree of variation in the environment. A monotonous environment tends to increase measured heritability.) There is also evidence from Western populations that genetic factors play a predominant role in observed variations in blood pressure and in susceptibility to diabetes, schizophrenia, manic-depressive psychosis, rheumatic fever, tuberculosis and bronchial asthma.

Other traits, of course, such as human body weight, have a low heritability, and variation in them is largely a function of the environment. Nevertheless, studies of heritability leave the impression that much of the observed diversity is genetic.

Discontinuous Variation

Weight and height vary continuously over an uninterrupted range of values and seem to be influenced by a great many genes. Certain other characters can have only a limited number of distinct states. The human blood groups are an example: a person can belong to group A, AB, B or O, but it is not possible to be in any intermediate position. For the purposes of genetic analysis such discontinuous variation has an important advantage: it is very often entirely ge-

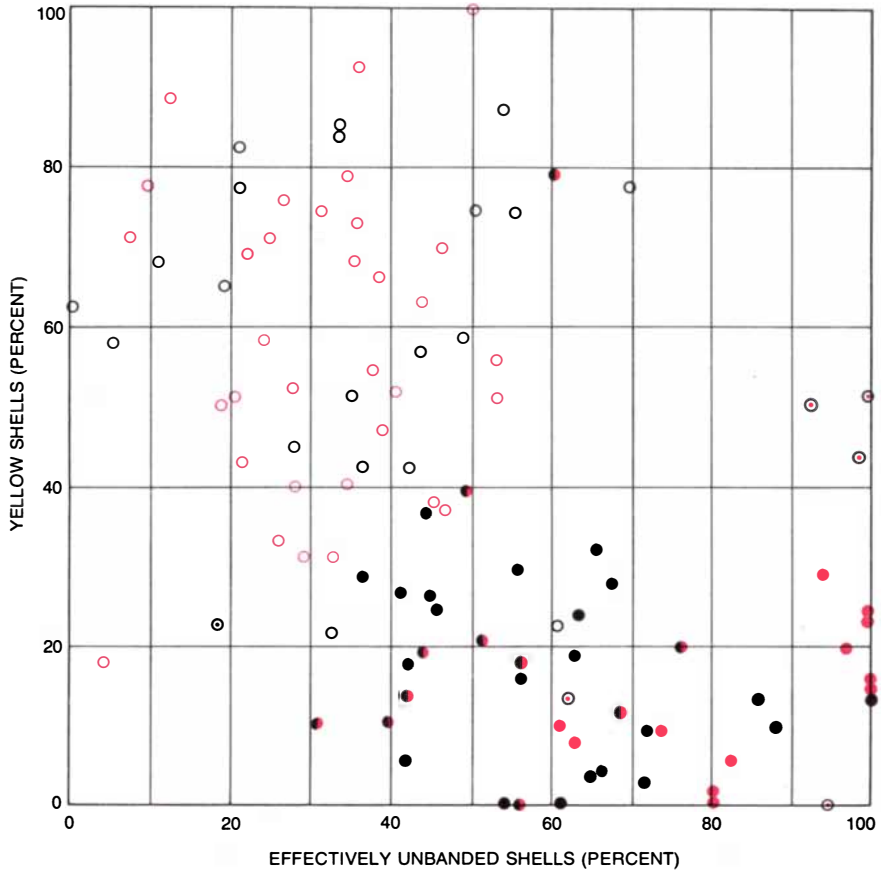
netic. However the environment is altered (short of causing death), a person with the allele for blood group A will display the antigens characteristic of group A. Furthermore, the inheritance of discontinuous variations is usually simple; in many cases they are controlled by one locus or a few loci.

The ABO antigens are just one of more than 30 known systems of blood groups, all of which are inherited in a relatively simple manner. There are many other discontinuous variations in man, such as the histocompatibility antigens, which determine the success or failure of organ transplantations. There are also a great many examples in other species.

One organism whose natural variation has been studied extensively is the land snail *Cepaea nemoralis*, common in Europe and in some areas of eastern North America. Among the observed varieties are snails with pink shells and snails with yellow shells; in addition shells of either color can be encircled by as many as five brown bands or stripes [*see illustration on opposite page*]. The color of the shell and the number of bands are controlled by genes at separate loci. There are many possible combinations of shell color and banding pattern, and there are several additional characteristics in which variation is found [*see top illustration on page 51*]. Shell color and banding are inherited in a straightforward manner, a fact that can be demonstrated by simple breeding experiments.

With J. J. Murray, Jr., of the University of Virginia, I have studied populations of another snail, *Partula suturalis*, which inhabits the island of Mooréa near Tahiti in French Polynesia. *P. suturalis* varies not only in pigmentation but also in the direction in which its shell is coiled. Some shells twist to the right and some to the left [*see bottom illustration on page 51*]. Again the pattern of inheritance is relatively simple, although the direction in which the shell coils is determined by the genes of the female parent rather than by those of the snail itself.

These examples of discontinuous variation share a curious property: the frequencies of the varieties are not those that would be expected from the classical Darwinian argument. In a population where mutations are usually disadvantageous and are quickly eliminated one would expect a single variety (the most successful one) to be common and any others to be very rare. The actual distribution of many discontinuously variable traits is quite different. Blood



CORRELATION OF SHELL MARKINGS with habitat in *C. nemoralis* implies that the patterns are influenced by natural selection. Each circle represents a population of snails; they were sampled by A. J. Cain and P. M. Sheppard of the University of Oxford. In hedgerows and rough herbage yellow, banded snails predominate; in woodlands yellow shells are less common and many shells are "effectively unbanded"; they lack two bands and appear to predators to be unbanded. If the markings were not subject to natural selection, the varieties would be distributed randomly.

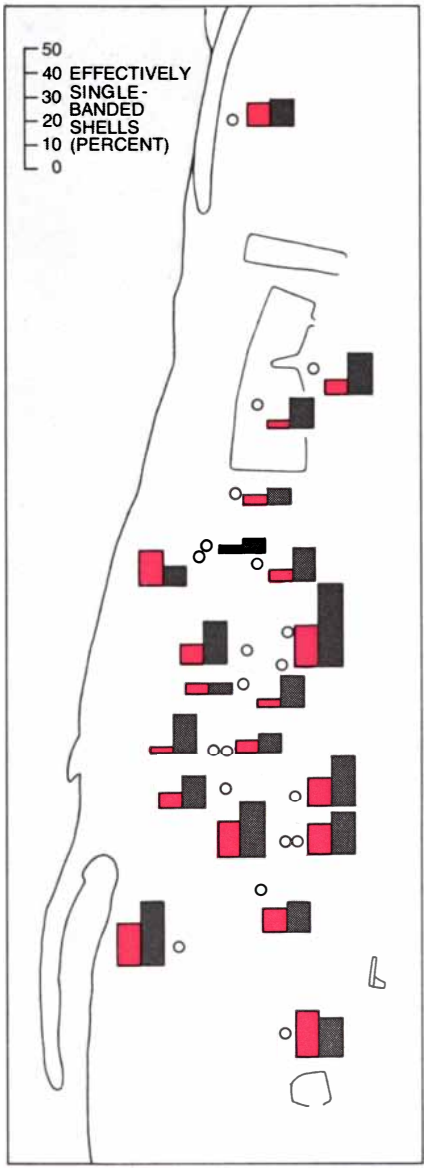
group O is probably the most common of the four ABO types, but the others are not rare; groups A and B may be equally common, and together with group AB they may constitute a large part of the human population. Similarly, in many populations of *C. nemoralis* the pink-shelled and yellow-shelled forms have approximately the same frequency.

Genetic Polymorphism

A population in which two or more distinct inherited varieties coexist at frequencies too great to be attributed to mutation is said to exhibit genetic polymorphism. Because it is not always clear what frequency is too great to be attributed to mutation, a population is now considered to be polymorphic only if the rarest variety has a frequency greater than 1 percent. In any large population that is certainly a greater frequency than could be maintained for long by a disadvantageous mutation.

Even if it is granted that genetic polymorphisms cannot represent disadvantageous mutations on their way to extinction, it might be argued that what is being observed are the rare occasions when advantageous mutations are spreading through a population. This process must take place in nature, and we should be able to observe it in progress, but it cannot account for the polymorphisms described here or for many others. An advantageous mutation should increase in frequency until the previously successful form is eradicated, so that the two forms should be found together only for a limited time. Several of our relatives among the apes, however, display a polymorphism resembling that of the human ABO blood groups, suggesting that that polymorphism was already established before the ape and

hominid lines diverged several million years ago. In the case of *C. nemoralis* fossil populations tens of thousands of years old exhibit the same polymorphisms for shell color and banding pattern as modern populations. If snails of one type are proliferating at the expense of the others, why was the process not completed long ago? Actually there is evidence that many polymorphisms are stable.



CHANGE IN FREQUENCY of the genes for shell banding was demonstrated by a long-term study of *C. nemoralis*. The snails were studied on a sand dune at Berrow in England, where increased vegetation has provided shelter for thrushes, which prey on the snails. The bars signify the percentage in each population of single-banded snails, a group included among the effectively unbanded snails. When surveyed in 1926 by Cyril Diver (*colored bars*), most of the populations had few single-banded shells; when reexamined in 1959 and 1960 by the author and J. J. Murray, Jr. (*black bars*), the proportion of shells having a single band had increased in almost all of the populations.

A further defense of classical Darwinism is possible. Even if it is conceded that certain traits are polymorphic in some populations, a crucial question is what proportion of all the genetic loci are polymorphic. If the proportion is small, the classical view remains substantially intact. Only if variation can be demonstrated at a fairly large number of loci must the classical view be revised.

The difficulty in answering this question lies in discovering how many genes do not vary. We have only crude guesses as to the total number of genes in a snail or a man, and those that are identical in all individuals leave no traces in the gross form of the organism by which they could be identified or counted. This problem has been circumvented by examining not the expression of the genes in the visible characteristics of the organism but the more nearly immediate products of the genes: the proteins they specify.

Proteins are made up of amino acids connected in the long chains called polypeptides. The properties of the protein are determined by the sequence of amino acids, and that sequence is in turn determined by the sequence of nucleotide bases in the DNA of the genes. In the process called transcription the double-strand DNA produces a single strand of RNA. In another process, called translation, the genetic program of the organism is "read" from the RNA in successive groups of three nucleotide bases (triplets). Each triplet specifies one of 20 amino acids, which is then attached to the end of the growing polypeptide chain. The completed polypeptide spontaneously assumes a characteristic three-dimensional form, and often it must be combined with other molecules to make a biologically active protein. It can then function as an enzyme or as a structural component or in other biological roles.

Through the selective manufacture of proteins genes influence development. The characteristics and behavior of an organism depend ultimately on the sequences of amino acids in its proteins, and evolution largely consists in the progressive substitution of one amino acid for another. Each of these changes reflects a corresponding change in a segment of the DNA molecule, but it should be noted that not all mutations in the DNA modify the protein. The genetic code is redundant, and some mutations merely produce synonyms.

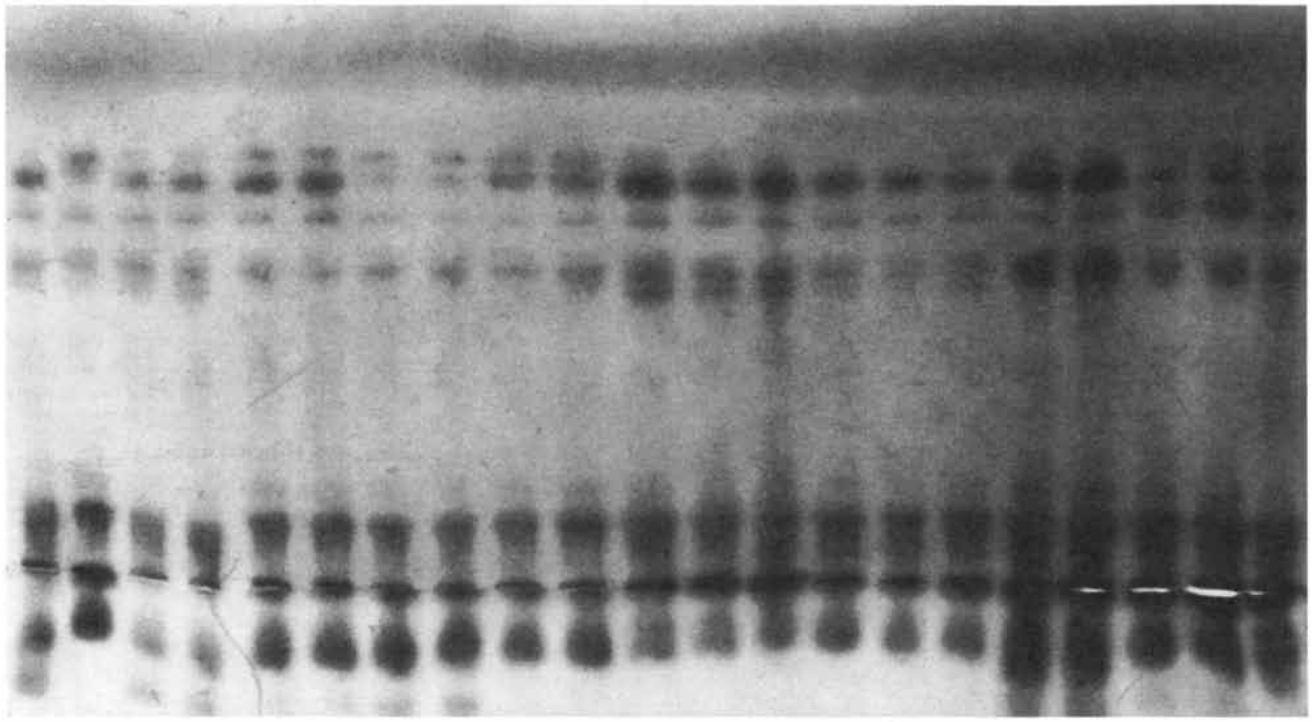
It is one of the triumphs of molecular biology that we can now determine the sequence of amino acids in proteins. For example, the protein cytochrome *c* is a single polypeptide composed of 104 amino acids; it is involved in the transport of electrons during cellular oxidation. Amino acid sequences have been determined for the cytochrome *c* of many species, ranging from yeasts to man [see "The Structure and History of an Ancient Protein," by Richard E. Dickerson; SCIENTIFIC AMERICAN, April, 1972]. Closely related species (such as man and monkeys) have similar sequences, whereas distantly related species (such as man and yeasts) have very different ones. By comparing the various sequences we can construct a table of resemblances that turns out to be very similar to the evolutionary trees derived from the more conventional methods of comparative anatomy and paleontology [*see illustration on page 56*]. Thus within the product of a single gene is contained information about the entire evolutionary history of an organism.

Gel Electrophoresis

The study of amino acid sequences has improved our understanding of variation between species; the same technique can be applied to the study of variation within species, but in order to do so it is necessary to examine proteins from many individuals. When such large samples are required, the determination of amino acid sequences is extremely laborious. There is, however, a simpler method of detecting amino acid substitutions; it is called gel electrophoresis. Like the analysis of amino acid sequences, electrophoresis enables us not only to identify those proteins that differ from individual to individual but also to detect those that are invariant. By means of electrophoresis we can therefore estimate the proportion of genes that are polymorphic.

Electrophoresis discriminates between variant forms of a protein on the basis of their electrical properties. Under biological conditions many amino acids carry an electric charge and hence the electrical properties of a polypeptide are influenced by its amino acid composition. In many cases (but not all) the substitution of one amino acid for another can alter those properties.

The alteration can be detected by subjecting the protein to an electric current. In practice proteins from several individuals are inserted into a gel made of starch or cellulose acetate or the



ELECTROPHORESIS of proteins reveals polymorphisms at the biochemical level in the snail *Partula taeniata*. Extracts from several snails were inserted into wells near the bottom of a gel made of the polymer acrylamide. When an electric current was passed through the gel, the proteins migrated at rates determined primarily by their amino acid composition. The gel was then stained to reveal the positions of a specific class of proteins: the enzymes called esterases. Each dark band on the gel represents an esterase. Some of the bands are common to all the specimens, indicating that the gene producing that esterase is identical in all individuals. Several bands, however, are present in some individuals but are absent in others, indicating that there is more than one allele, or form of the gene, in the population. By counting the bands that vary and those that do not one can estimate proportion of polymorphic genes.

synthetic polymer acrylamide. When an electric current is passed through the gel, the speed with which the protein migrates is determined by the electric charges of its amino acids. (The size and conformation of the protein can also have an effect.) Two proteins can be compared by inserting them next to each other in the gel and seeing how far they move in a particular interval. Their position at the end of the period is revealed by applying a stain specific for the protein being studied; when the protein is an enzyme, the stain is often coupled to some product of the reaction catalyzed by the enzyme. Electrophoresis can distinguish proteins that differ by only one amino acid.

We can discover what proportion of genes are polymorphic by selecting specimens from a population and testing a number of proteins from each of them by electrophoresis. The result will reveal which proteins have different mobilities in different individuals and which are the same in all members of the sample population. Because each polypeptide is the product of a single gene, the technique enables us to estimate how many genes have several alleles in the population and how many have only one.

Michael S. Johnson in the author's laboratory at the University of Nottingham has investigated protein polymorphisms in the snail *Partula taeniata*. Several snails were homogenized, and the extracts were inserted at the bottom of an electrophoretic gel. Current was then passed through the gel, and it was stained to reveal the activity of a particular enzyme or a group of related enzymes. In one case, for example, the stain identified esterases (enzymes that mediate the formation and breakdown of esters). The extract from each snail showed a number of bands, each containing an esterase; some of the bands were the same in all the snails, but others varied from one snail to another [*see illustration above*]. By counting the bands that vary and the bands that do not we can estimate the number of esterase genes that vary and the number that do not. If we repeat the procedure for a variety of enzymes, and if we can assume that they represent a random sample of all genes, we can estimate what proportion of the genes are polymorphic.

Protein polymorphisms were first systematically examined in man by Harry Harris of University College London and in fruit flies by Richard C. Lewontin and John L. Hubby of the University of Chicago. They concurred in finding that about 30 percent of the genes that specify enzymes are polymorphic for electrophoretic mobility.

It is likely that electrophoresis both overestimates the number of invariant genes and underestimates the number of variable ones. A single protein can change its conformation in the electrophoretic gel and produce two or more bands; each band might then be interpreted as a separate, invariant protein, thus inflating the estimate of the number of invariant genes. At the same time many proteins that do differ in composition cannot be distinguished by electrophoresis because not all amino acid substitutions alter the electrical properties of the molecule. The net effect of these inaccuracies is a tendency to underestimate the proportion of polymorphic genes; it may be that half or more of the genes for enzymes are polymorphic. An additional problem is that the enzymes examined almost certainly do not represent a truly random sample of all genes. Nevertheless, a wide range of proteins have now been studied, from a wide range of organisms including plants as well as animals, and the estimates from

the various species are reasonably consistent. (The degree of polymorphism does seem to depend, however, on the breeding habits of the species; less polymorphism is sometimes observed where inbreeding is common.)

Evidence for widespread protein polymorphism is fatal to the classical interpretation of variation. Natural populations are not in general composed of one common genetic type with a few rare mutants. On the contrary, they are extremely diverse.

Selection for Variation

How can we explain this extraordinary degree of variation? Two categories of explanation have been proposed. One group of geneticists has suggested that the diversity is present because it has no influence on the fate of the organism. They argue that the substitution of one amino acid for another in a protein made up of hundreds of amino acids is unlikely to affect the properties of an enzyme. The alleles responsible for such substitutions can spread through a population by random processes. They are neutral with regard to natural selection, conferring neither advantage nor disadvantage. In short, they are evolutionary noise.

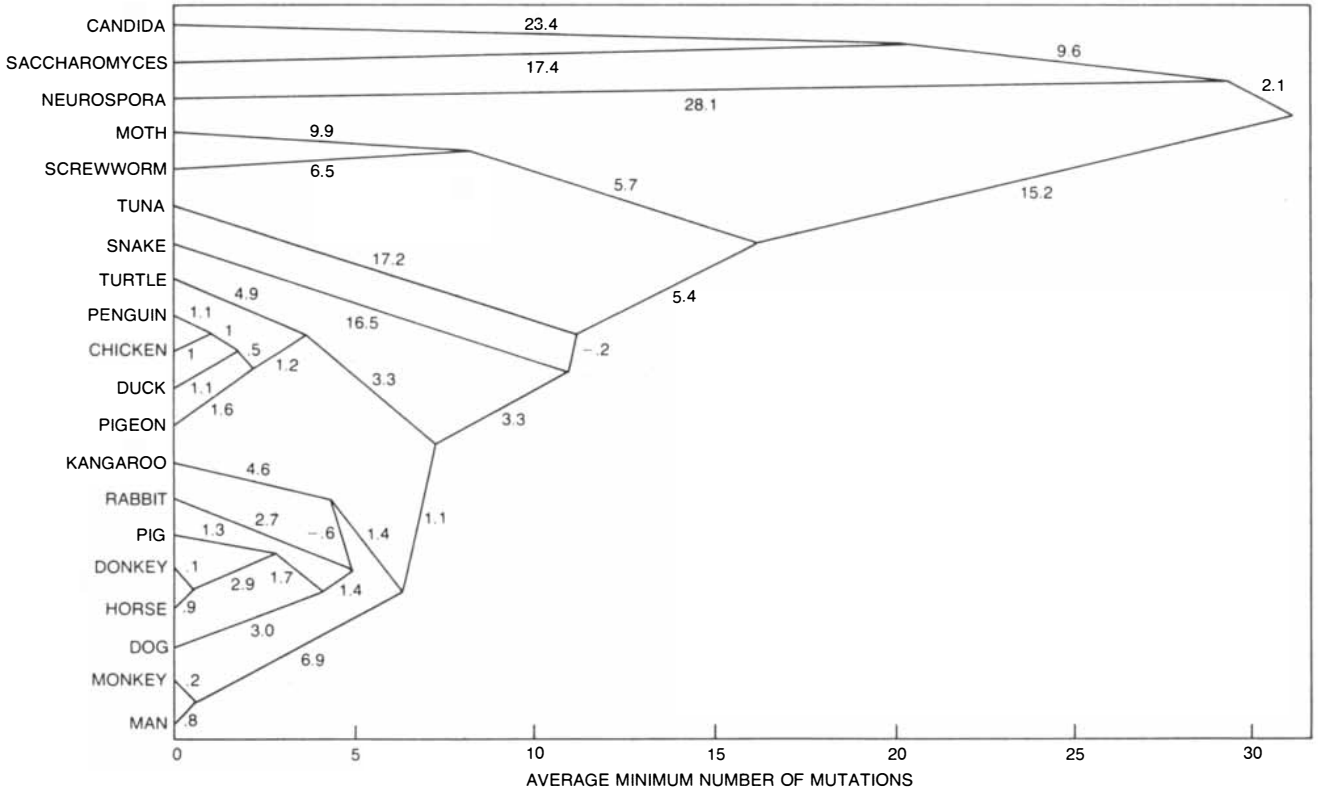
Mathematical models have been devised to show that the amount of variation observed today and the changes that have taken place during the evolution of proteins can be explained in terms of alleles that do not affect the survival or reproduction of the individuals carrying them. This "neutralist" view has appealed to geneticists of a mathematical bent, particularly in the U.S. and Japan. It rescues the classical argument by conceding that polymorphism is widespread but asserting that it is irrelevant.

Other workers, with a more direct interest in the ecology of the organisms, have been impressed by the fineness of their adaptation to the environment. They argue that diversity has developed because natural selection has favored it. There are two components to this "selectionist" argument: first, that the polymorphic genes do affect the survival and reproduction of the individuals carrying them, and second, that widespread systems in nature actively maintain genetic diversity.

We can, in specific instances, discriminate experimentally between the neutralist and the selectionist hypotheses. The results strongly suggest that natural selection acts to maintain polymorphisms.

A number of known polymorphisms can be related to factors that influence fitness. Some of the human blood groups, for example, are associated with differential susceptibility to disease. (The blood groups are not themselves protein polymorphisms that can be detected by electrophoresis, but they are the products of genetic variations in the activities of enzymes that are proteins.) People of blood group O are apparently more likely than others to contract the A_2, or "Asian," form of influenza. The histocompatibility antigens are associated with differing susceptibilities to various other diseases, including rubella, multiple sclerosis and allergies such as bronchial asthma.

The shell-color polymorphism in *C. nemoralis* is also subject to natural selec-



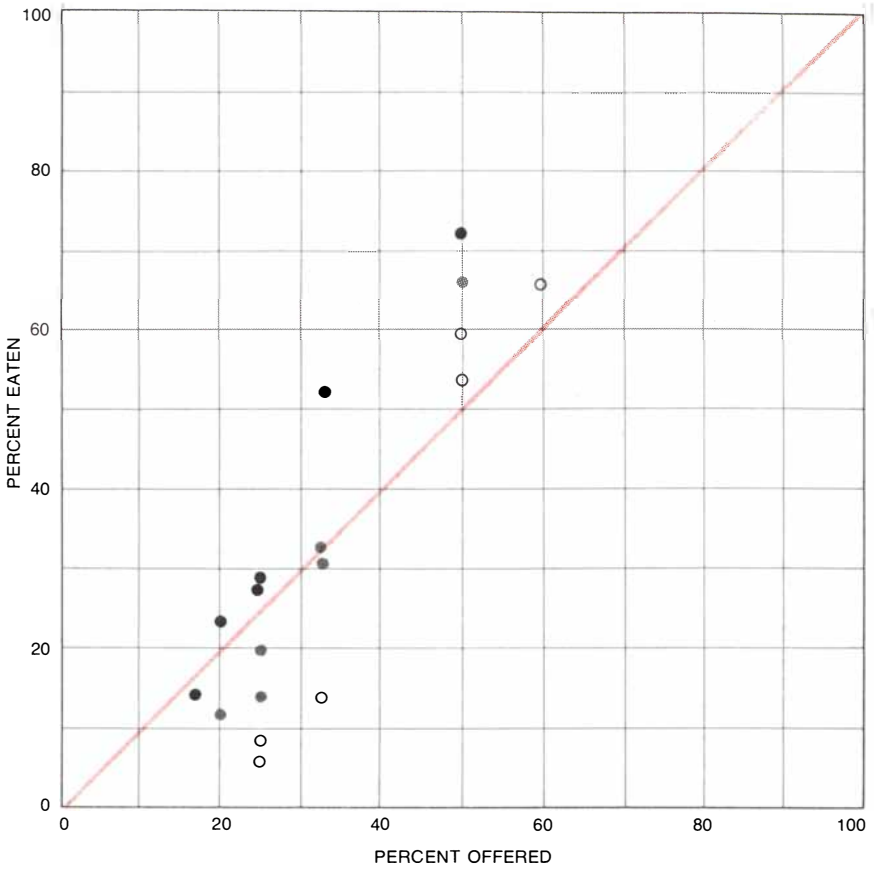
EVOLUTIONARY HISTORY of a protein was deduced from a comparison of amino acid sequences in various species. The protein is cytochrome *c*, which has 104 amino acid units. The substitution of one amino acid for another represents at least one mutation in the gene that codes for the protein. Numbers on each line segment give the minimum number of mutations separating points of divergence. Vertexes are placed at the average distance calculated for all lines of descent. Vertebrates and invertebrates are separated by at least 15 mutations in the cytochrome *c* gene, and plants and animals are separated by more than 30. The evolutionary tree was devised by Walter M. Fitch of the University of Wisconsin Medical School and Emanuel Margoliash of Northwestern University.

tion. Populations of the snail in the region of Oxford in England have been studied by A. J. Cain and P. M. Sheppard of the University of Oxford. They have found that in beech woods there are relatively few yellow shells and a disproportionately large number of shells that are effectively unbanded, or missing the top two bands; such shells appear from above as if they were entirely unbanded. In grasslands the proportions are reversed: yellow shells are relatively common and effectively unbanded ones are relatively rare. This correlation of pattern with habitat would not be expected if the genes for shell color and banding were neutral; in that case the frequency of each type would vary randomly.

There is, in fact, direct evidence that natural selection acts on the shell markings of *C. nemoralis.* Among the important predators of the snails are thrushes that capture them and break the shells on stones. The broken shells can be collected and compared with the population from which they were taken. Sheppard has shown by this method that the thrushes preferentially select the most conspicuous snails. Clearly in this case the color of a snail's shell affects its probability of survival.

A long-term study of several populations of *C. nemoralis* on a sand dune at Berrow in England has demonstrated that differences in survival between the color types can gradually change the genetic composition of a population. The ecology of the dune has changed as a result of the invasion of a shrub, the sea buckthorn, which has apparently provided shelter for increasing numbers of thrushes. The snail populations were first surveyed in 1926 by the late Cyril Diver; in 1959 and 1960 Murray and I reexamined the same populations. During the interval between the surveys (approximately 12 snail generations) the frequency of shells with only a single band increased significantly [*see illustration on page 54*]. The consistent change in many separate populations could only have been a result of natural selection.

It is well known that in industrial areas the frequencies of different wing colors in certain moths have changed in a similar way. Species of moth that are polymorphic for dark or light coloration have responded to industrial air pollution, which blackens the trees on which the moths rest, by an increase in the frequency of dark forms. In some areas of England where airborne soot has recently been reduced the frequency of dark-winged moths is now decreasing [see "Moths, Melanism and Clean Air," by J. A. Bishop and Laurence M. Cook; SCIENTIFIC AMERICAN, January].



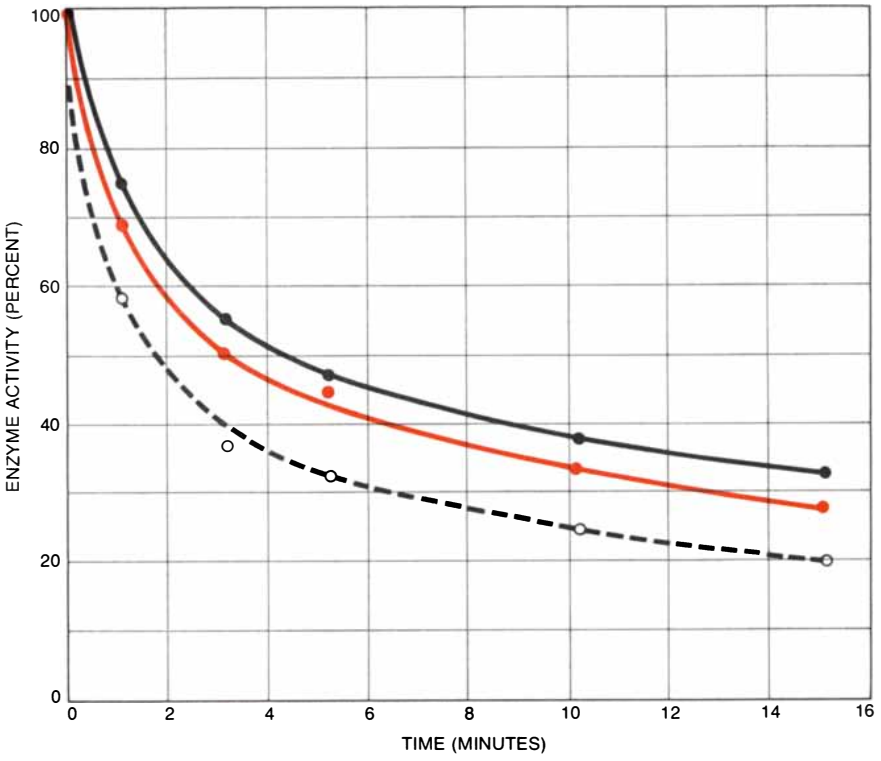
FREQUENCY-DEPENDENT SELECTION was demonstrated by E. J. Popham with water bugs polymorphic for color. The water bugs (of the species *Sigara distincta*) were fed to minnows in groups containing various proportions of pale (*open dots*), medium brown (*gray dots*) and dark brown (*black dots*) forms. If the minnows had exercised no selection, each type would have been eaten in the same proportion as it was supplied; actually the rarer forms were eaten less often than was expected. Through this mechanism a trait becomes advantageous only when it is rare and polymorphism is maintained in the population.

These examples establish that some polymorphic genes do significantly affect fitness. One might reasonably object, however, that they are not representative, and in particular that they do not represent the kind of variation detected by protein electrophoresis. That objection can be answered only by examining the proteins themselves. As yet few polymorphic proteins have been thoroughly studied. Nevertheless, several variants that must differ in fitness have been discovered.

Maintenance of Diversity

One of the proteins that is polymorphic in human populations is glucose-6-phosphate dehydrogenase, an enzyme that promotes one of the initial reactions in the oxidation of glucose. A variant form common in some parts of the world, which is believed to differ from the predominant form by a single amino acid, apparently confers increased resistance to malaria. The malaria parasite (a protozoon of the genus *Plasmodium*) thrives in red blood cells of people with one type of glucose-6-phosphate dehydrogenase but not in those of people with the other type. The variation is evidently important to survival.

If we suppose for the moment that all polymorphic variants are affected by natural selection, it is still necessary to explain why in each case one of the variants does not replace all the others. Why does polymorphism persist? Why has the form of glucose-6-phosphate dehydrogenase that increases resistance to malaria not eliminated the other form from the population? The persistence of such polymorphisms can be explained only by postulating widespread systems that actively maintain diversity in spite of the selective forces tending to eliminate it. These systems must include feedback mechanisms capable of restoring equilib-



TEMPERATURE SENSITIVITY of the enzyme alcohol dehydrogenase was measured by the rate at which its activity declines at 40 degrees Celsius. In populations of the fruit fly *Drosophila melanogaster* there are two alleles of the gene that specifies the enzyme; the variant forms can be distinguished by electrophoresis and are designated *S* and *F*. The activity of alcohol dehydrogenase extracted from flies homozygous for the *S* gene (*solid black line*) declines more slowly than that from flies homozygous for the *F* gene (*colored line*); the enzyme of heterozygotes (*broken black line*) apparently loses its activity fastest of all. Differences in heat sensitivity affect metabolism and should be subject to natural selection.

rium when the system has been disturbed. Several possible mechanisms are known, of which two seem likely to be the most important.

The first is called heterozygous advantage. It depends on the fact that in higher plants and animals every adult has two sets of chromosomes and therefore two genes at each chromosomal locus. As we have seen, the population at large can contain several alleles at a locus, but an individual can carry no more than two of them. If he carries two identical genes, he is homozygous; if he carries two different ones, he is heterozygous. When heterozygotes survive or reproduce better than homozygotes, both alleles can be actively maintained in the population.

A now familiar example of heterozygous advantage has been discovered in tropical and subtropical Africa, where an allele prevalent in the human population produces a variant form of the protein hemoglobin. The variant hemoglobin differs from the normal form by a single amino acid and can be distinguished from it by electrophoresis. In people homozygous for the anomalous hemoglobin the red blood cells tend to collapse, assuming a form that was originally described as sickle-shaped; the result is a severe hemolytic anemia that is usually fatal before reproductive age. Since the gene is obviously disadvantageous, why does it persist in the population? The answer is that heterozygotes (who carry both the normal and the abnormal allele) are exceptionally resistant to malaria. In parts of Africa where malaria is a major cause of mortality people homozygous for the normal gene tend to die of malaria and those homozygous for the sickle-cell gene tend to die of anemia. The heterozygotes survive and the polymorphic genes persist in equilibrium.

It is possible that many polymorphisms are maintained by heterozygous advantage. It has been argued, however, that all of them cannot be, since each instance of heterozygous advantage demands an increased mortality (or reduced fertility) for both homozygotes. If a great many genes were subject to this kind of selection, almost every individual would be homozygous for at least one of them, and the burden of mortality and infertility would be too great for the population to bear. Whether or not this argument is valid, there is another mechanism for maintaining polymorphism. It is called frequency-dependent selection.

In the 1940's E. J. Popham of the University of Manchester studied predation on water bugs by minnows. He placed water bugs of various shades of brown in a fish tank with a uniform background of sand and recorded how many insects of each color were eaten. In one series of experiments different proportions of bugs of various colors were put in the tank. In the absence of any selection one would expect the bugs of each color to be eaten in the same proportions in which they were supplied. Actually Popham found that when a color was rare, insects of that color were eaten less often than expected, and when it was common, they were eaten more often. In other words, the advantage or disadvantage conferred by a variant was found to depend on the frequency of the variant [*see illustration on preceding page*]. Such a mechanism could maintain several colors in the population.

Lukas Tinbergen, Harvey Croze, John Allen, Ian D. Soane, David T. Horsley and I have found that frequency-dependent selection is often encountered in the behavior of hunting animals. Not only fish but also predatory birds and mammals seem to concentrate on common types of prey and to overlook rarer types. Birds hunting by sight should therefore maintain color polymorphisms in their prey populations (as in the snail *C. nemoralis*), and mammals hunting by smell should maintain olfactory polymorphisms. Disease organisms may behave in a similar way, becoming adapted to the biochemical environment of their commonest host. If they do, biochemical variants of the host will retain a selective advantage as long as they are rare; when they become common, the microorganism will gradually become adapted to the variant form. Through the interaction of host and pathogen the biochemical polymorphisms could be maintained.

Some protein polymorphisms are apparently subject to frequency-dependent selection, as is indicated by the experiments of Ken-Ichi Kojima and Y. N. Tobari of the University of Texas and by work in my laboratory. In wild populations of the fruit fly *Drosophila melanogaster* there are two common varieties of the enzyme alcohol dehydrogenase, which converts alcohols to aldehydes or ketones. The two varieties can be distinguished by electrophoresis. Kojima and Tobari arranged the crossbreeding of flies so that varying proportions of

eggs homozygous or heterozygous for the alcohol dehydrogenase gene were introduced into a limited habitat inside a glass vial; then they recorded what proportion of each type survived to adulthood. Each variant survived better when it was rare than when it was common. Our own experiments, in which eggs or very young larvae were counted into the vials, confirmed these results.

In these cases frequency-dependent selection could not have been brought about by the action of predators, since there were none in the vials. It might possibly have been produced by disease organisms. It seems far more likely, however, that it reflected the abilities of flies with differing enzymes to exploit different components of the environment. When a fly with one kind of enzyme was rare, the component it could most effectively utilize was relatively in excess; when the fly was the commoner variety, the environment became relatively deficient in that component. As a result of this effect a polymorphic population should exploit its habitat more efficiently, and hence should attain larger numbers, than a genetically uniform population.

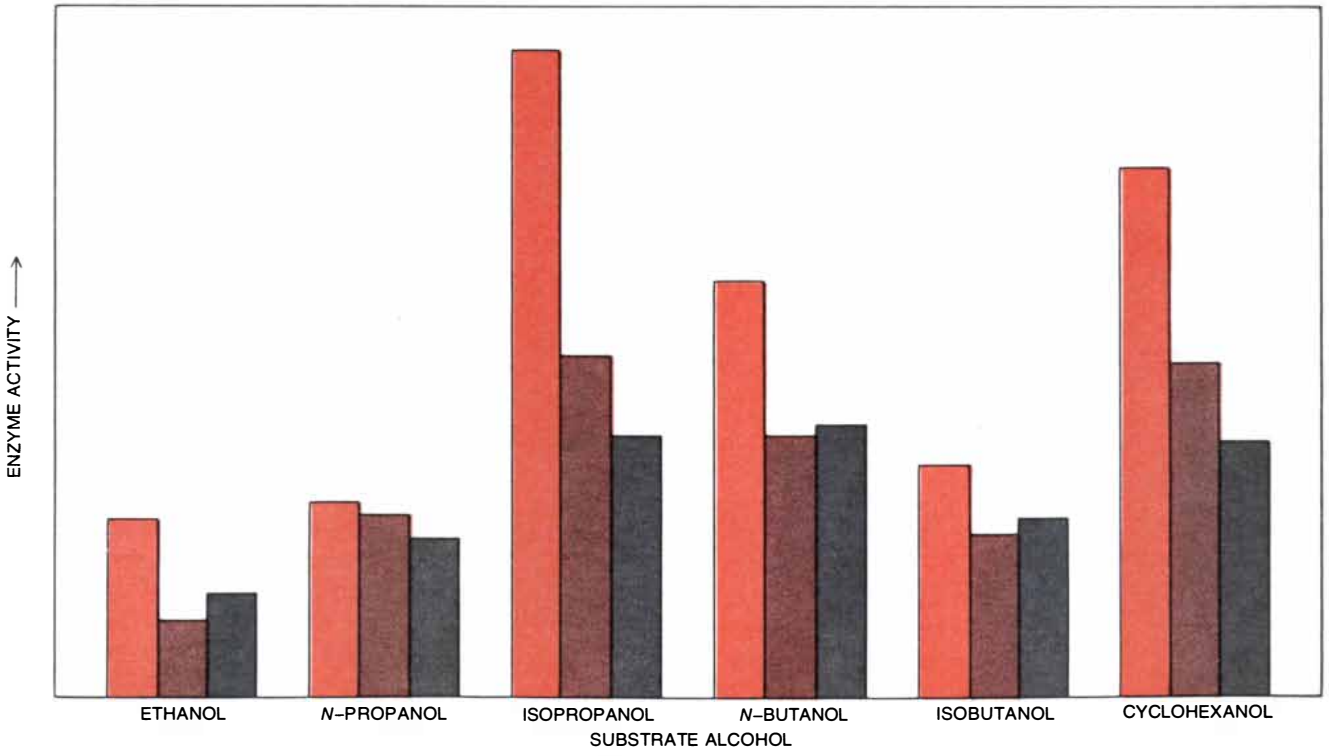
The recent work on *Drosophila* alcohol dehydrogenase provides strong evidence that an enzyme polymorphism can be maintained by natural selection. One argument remains, however, in defense of the neutralist hypothesis. It is possible that natural selection acts not on the alcohol dehydrogenases but on the products of some other gene close to the alcohol dehydrogenase gene on the same chromosome. Thus one could assume that a few polymorphic genes scattered along the chromosome are subject to natural selection, but that the majority are neutral and are merely carried along for the ride.

Enzyme Effectiveness

This hypothesis can be refuted only by direct biochemical studies of the proteins. It is necessary to show that they differ not only in electrophoretic mobility but also in other ways more likely to be of significance in the economy of the organism. In our laboratory at Nottingham, Thomas H. Day, David T. Parkin, Terence O. Dobson, Philip C. Hillier, Phillip D. Morgan and Lesley Needham have undertaken such a study. Our first results cast serious doubt on the neutralist hypothesis.

We examined the *D. melanogaster* alcohol dehydrogenase, labeling the variants *F* and *S* (for "fast" and "slow" in electrophoretic mobility). We found that on a substrate of ethanol the *F* enzyme is almost twice as active as the *S*. On the other hand, the *F* protein is less stable at high temperature, and its activity declines more rapidly than that of the *S* protein. The two forms of the enzyme also differ in their ability to act on different substrate alcohols: the *F* form is only slightly more active than the *S* with *n*-propanol, but it is more than twice as active with isopropanol. Finally, acidity and alkalinity affect the two variants differently.

It seems most unlikely that these large disparities in enzyme activity and specificity could be unimportant to an organism that spends its youth in rotting fruit, a veritable bath of alcohol. Indeed, we now have evidence that the alcohol dehydrogenase variants are directly subject to selection. Morgan has shown, by introducing a variety of alcohols into the food of fruit flies, that the intensity of selection in favor of the *F* form is strongly correlated with the ability of that variant to metabolize the alcohol. Furthermore, penten-3-ol, an alcohol converted by the enzyme into a toxic ketone instead of an aldehyde, selects for the less active *S* variant. C. L. Vigue, Franklin M. Johnson and A. Powell of North Carolina State University at Raleigh have demonstrated that heat shock (at 40 degrees Celsius) also favors the *S* form. In



ENZYME ACTIVITY of polymorphic variants of alcohol dehydrogenase depends on the nature of the substrate alcohol. For six alcohols tested the enzyme of the *F* homozygote (*color*) is more efficient than that of the *S* homozygote (*gray*) or the heterozygote (*gray and color*). Such variations would be expected to give a competitive advantage to different genotypes in different environments.

the wild populations of *D. melanogaster* in the eastern U.S. the *S* gene becomes commoner as one moves south. These observations are precisely what would be expected if the alcohol dehydrogenase polymorphism were directly influenced by selection.

The neutralist hypothesis can now be defended only by asserting that these findings are exceptional and that the cases of hemoglobin, glucose-6-phosphate dehydrogenase and alcohol dehydrogenase are not representative of protein polymorphisms in general. Similar results are now being obtained, however, in studies of other proteins, including the amylases of fruit flies and the human histocompatibility antigens. Harris has reported that of 23 human protein polymorphisms studied 16 manifest biochemical differences, apart from electrophoretic mobility, between the allelic variants. Although it remains to be shown that these differences are detected by natural selection, the lesson of alcohol dehydrogenase suggests that selective mechanisms will be found.

The case for the classical and the neutralist views of variation is now very weak. It has been demonstrated that most natural populations of plants and animals are genetically heterogeneous. Moreover, there is strong evidence that the diversity of forms exists because natural selection favors it, that is, because the variants themselves affect the survival and reproduction of the individuals carrying them.

Consequences of Theory

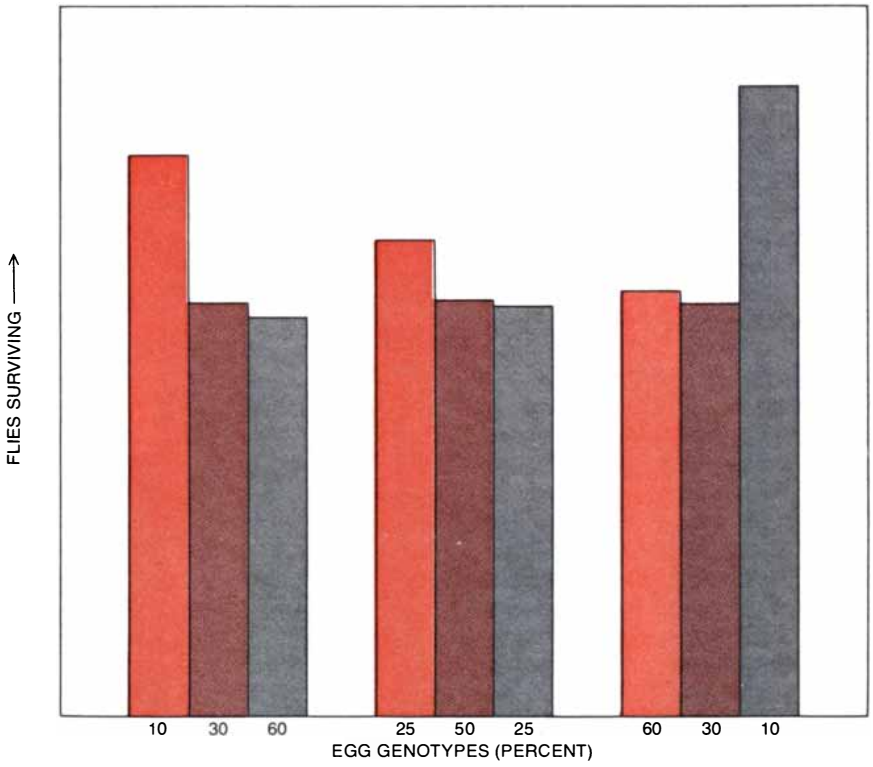
If this conclusion is correct, it has several important implications. To begin with, it requires that we revise our notion of how evolution proceeds. Evolution, it now appears, is not an orderly progression in which one type replaces another but a complex flux of shifting dynamic equilibriums. In applied biology our conclusions suggest that many agricultural breeding programs may be misdirected. If a genetically diverse population can better exploit its environment, we ought not to be breeding genetically uniform crops. This principle is likely to prove particularly important in less developed countries, where the agricultural environment is often quite heterogeneous.

In the realm of human biology our conclusions imply that people differ genetically far more than was once thought. Two people chosen at random from the American or the British population will differ at hundreds of chromosomal loci, and possibly at thousands. These differences are not trivial: they affect, or in the past have affected, our survival and reproduction. Many of them are associated with differing susceptibilities to disease, and if those associations can be analyzed, medicine will be able to take account of them.

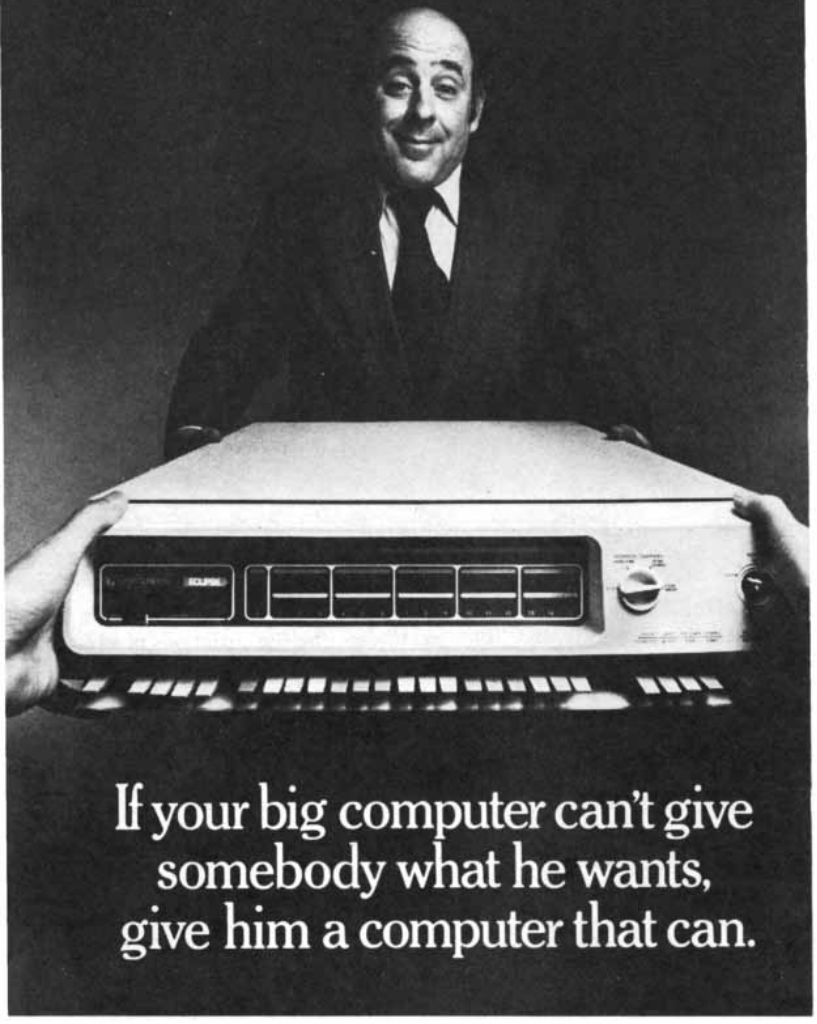
Already investigators at Nottingham and at Stanford University are studying the genetics of susceptibility to schizophrenia. The method involves searching for polymorphisms among the enzymes controlling the metabolism of aromatic amino acids, since there is reason to believe those amino acids are involved in the etiology of schizophrenia. Human polymorphism is also the subject of the new medical science of pharmacogenetics, which investigates differences between individual reactions to therapeutic drugs. It is clearly a mistake to suppose all people will react in the same way to a drug when so many of their enzymes are manifestly different.

The impact of Darwin's theory of evolution on areas of thought outside biology is well known. Following an era in which the Bible was read as a biological and geological treatise, *Origin of Species* in its turn became a treatise on religion and ethics and ultimately on politics and sociology. The modification of the theory in order to accommodate widespread polymorphism could produce as profound a revolution in the social sciences as Darwin's original statement did. It seems most unlikely, for example, that the observed diversity in our biochemistry could fail to have an effect on our behavior. The existence of this diversity might encourage the search not for the ideal social or political system but for the ideal array of social and political systems. We should, perhaps, ask for polymorphism in our institutions to match the polymorphism in ourselves.

In conclusion, I should like to complete the quotation with which I began this essay, and leave the last word to Sir Thomas Browne: "It is the common wonder of all men, how among so many millions of faces, there should be none alike: now contrary, I wonder as much how there should be any. He that shall consider how many thousand several words have been carelessly and without study composed out of 24 letters; withal how many hundred lines there are to be drawn in the Fabrick of one man, shall find that this variety is necessary."



POPULATIONS OF FRUIT FLIES polymorphic for the alcohol dehydrogenase gene are apparently subject to frequency-dependent selection. Eggs homozygous for the *S* gene (*gray*), eggs homozygous for the *F* gene (*color*) and heterozygous eggs (*gray and color*) were placed in vials with a fixed supply of nutrients. When the flies reached maturity, the proportion of each genotype was dependent on its initial frequency; rare genotypes were more successful than common ones. The advantage apparently conferred by rarity can be explained by the fact that nutrients consumed by common genotypes were depleted faster.



If your big computer can't give somebody what he wants, give him a computer that can.

It's hard for one computer to be all things to all people.

So if your big computer can't do what somebody wants done, get a small computer to do the job.

A Data General computer.

At Lowe's Companies (a group of 130 building material stores) the central computer knew the current prices and stock levels of the 7,000 items in stock at each store. But the salesmen didn't.

So Lowe's data processing management put Data General computers in their stores to give the salesmen all the information they needed. Our computers cost a third less than the IBM 370 it would have taken to do the job. And ours do the job better.

At the end of every day, our computers automatically report to the computer center in the home office. And by the beginning of the next day, the central computer sends each store the latest prices and new stock information.

Data General computers not only make order-taking easier and more accurate, they also cut the time needed to take an order by 30% and reduce inventory carrying costs, pricing errors, and bad debt losses.

RCA Records put our computers in the hands of their operating management so they could do order processing and inventory control without getting caught up in the complications of a big computer.

One of the largest banks in the Northeas put a Data General computer in their trust department to give their portfolio officers access to vital investment information in their large IBM computer. And our computer does the job for much less than it cost for the large computer to do it alone.

So instead of giving people reasons why your big computer can't do something, write for our brochure, "The Sensible Way to Use Computers."

Then when somebody starts complaining about your computer, you'll know enough to tell him where to go.

SCI–5

NAME

TITLE

FIRM

ADDRESS

CITY STATE ZIP

Data General, Dept. G, Route 9, Southboro, Mass. 01772 (617) 485-9100. Data General (Canada) Ltd., Ontario, Data General Europe, 15 Rue Le Sueur, Paris 75116, France. Data General Australia, Melbourne (03) 82-1361/Sydney (02) 908-1366

Data General

THE PERCEPTION OF SURFACE COLOR

Colors are normally regarded in terms of wavelength and intensity. The colors of surfaces are much more complex. Consider the subtle variations of color in a blue scarf carelessly thrown on a table

by Jacob Beck

View a wall some distance away through a hole in a dark screen, and the part of the wall within the hole will no longer be perceived as a material substance, standing in a specific illumination and having a definite spatial location. Since it is possible to focus one's eyes only on the edge of the hole and on the screen, the texture of the distant wall will be blurred, and an area of uniform luminance and hue will be projected on the retina. The color seen through the hole will appear as a filmlike expanse rather than as an attribute of a surface. Homogeneous areas of luminance and hue in a dark surround lead to what is termed a nonsurface perception of color.

Colors viewed in this way, sometimes called film colors, vary in brightness, saturation and hue. Brightness is the attribute that can be described as ranging from dim to dazzling. Saturation is the attribute that determines the degree to which a color differs from a white of the same brightness. Hue is the attribute that makes it possible to classify colors as red, yellow, green, blue or intermediates of these. Geometrically the family of film colors can be described in terms of a double cone. Variations in brightness are represented along the axis of the cone. Variations in hue are represented around the circumference. Variations in saturation are represented along the radii. Colors of increasing saturation are located at increasing distances from the axis. The highest degree of saturation is perceived for colors of medium brightness; as brightness increases or decreases, the attainable saturation becomes less. The attributes of hue, saturation and brightness exhaust the degrees of freedom of a film color. If two film colors are separately equated for their hue, saturation and brightness, the colors will appear identical.

The term surface color is used to refer to a color that is perceived as an attribute of a surface. The simplest condition for the perception of a surface color is to present a stimulus consisting of two different luminances adjacent to each other. If an orange light of homogeneous luminance and hue is viewed through a hole in a white screen that is at a slightly higher luminance, for example, the color within the hole will be seen not as a film color but as the color of an opaque surface. The color will be seen as though it were a piece of orange paper pasted on the white screen.

Surface colors exhibit attributes not possessed by film colors. Thus there are no gray, brown, maroon or olive green film colors. These colors represent variations in lightness that emerge when one views contrasting luminance variations, such as those that typically give rise to the perception of a surface. Physically the difference between a yellow and a brown is not the same as that between, say, a yellow and a red or a green and a blue. Differences in hue correspond to differences in wavelength. The wavelength composition of the light reflected from a yellow surface and from a brown surface can be identical. Differences in lightness do not correspond to differences in wavelength. The perception of brown can accordingly result from a contrast effect in which an adjacent area of higher luminance darkens the yellow color. The double-cone illustration can also be used to describe the variations of surface colors. Variations in lightness from white through gray to black are now arranged from top to bottom along the axis of the cone. Variations in hue and saturation are represented as they are in the case of film colors [*see illustration on page 66*].

The attributes of hue, saturation and lightness do not exhaust the variations possible with surface colors. Two surface colors that have been equated on the basis of these three attributes will not necessarily appear identical. For example, the appearance of a surface color changes with its illumination. If a white card is folded in half and placed so that one side of the card is illuminated and the other side is in shadow, both sides of the card are seen as white [*see top illustration on page 68*]. Yet the two sides do not look the same. One is able to perceive the difference in illumination. The illuminated side is seen as a bright or strongly illuminated white, the shadowed side as a dull or dimly illuminated white. The ability of the human perceptual system to separate the lightness of the surface from the brightness of the illumination is a manifestation of what psychologists refer to as color constancy. This term is used to describe the tendency of colors to retain their daylight appearance in spite of large changes in the

VARIATIONS OF SURFACE COLORS are captured in the still-life photograph on the opposite page. In spite of large differences in the intensity and wavelength of the illumination the human visual system tends to separate the perceived hue and lightness of a colored surface from the hue and brightness of the illumination; psychologists refer to this tendency of surface colors to retain their normal daylight appearance as color constancy. In addition to the various surface colors represented, which result primarily from differences in the composition, texture, smoothness and depth of finish of the assorted objects, an example of a film color is included; it is the red surface seen through a hole in black cardboard.

Pictures you see in seconds can save you hours.

In scientific detection.

When photography is used for scientific detection, special lighting and filtration is often required. Just coming up with the right combination using conventional methods (shooting, developing, adjusting light and filters, and reshooting) can take hours, even days.

Fortunately, there's a faster, more economical way—by using Polaroid's Type 105 or Type 55 positive/negative films on standard view cameras. With Type 105 (3¼ x 4¼ inch pack format) and Type 55 (4 x 5 inch sheet format), seconds after the picture has been taken you can evaluate the results, make adjustments and reshoot. Until it's just right.

And since both films provide a high-resolution negative along with the positive print, enlargements can be made for closer scrutiny, or multiple copies for distribution.

The picture on the left is an ultraviolet reflectogram that tells the art conservationist that the painting in question has badly deteriorated be-

neath the surface. The deterioration took decades. The proof took seconds.

In record keeping.

How can a laboratory assure that it has accurate visual records of thousands of experimental results? One way is to use a central photo lab,

MP-4 MULTIPURPOSE CAMERA

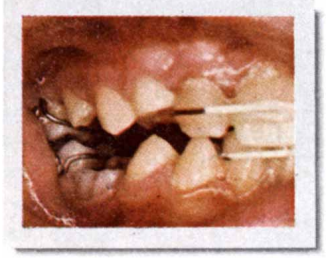
which is costly, time consuming and sometimes impossible.

A better way is to take pictures with a centrally-located Polaroid MP-4 Multipurpose Land camera. It can be operated by just about anyone with a little training. No darkroom is required. And there are 15 different Polaroid self-developing films (from black and white slides to 4x5 color prints) to choose from. Best of all, the pictures are available instantly — to clip into a lab notebook, to be routed, or to be filed.

The picture above, for example, was one of a series taken on our brilliant new Polacolor 2 film. It illustrates the effect of a specific amount of ozone on a tobacco cultivar.

In making a diagnosis.

This is Jeffrey's malocclusion. A year from now, it will be considerably improved. And pictures taken at every stage will tell just how the treatment is progressing.



CU-5 CLOSE-UP CAMERA

Many dentists now make such pictures on the spot to study, then to show at consultation meetings and later to file for reference — using Polaroid's hand-held CU-5 Close-up Land camera. It requires virtually no experience to operate and there's little or no guesswork involved.

The CU-5 can be used effectively anywhere that same-size, 2 : 1 and 3 : 1 magnification photographs are needed in color or black and white. Everywhere from patrol cars to computer readout terminals. Perhaps you have an application we haven't thought of.

In rapid communication.

How can you describe the need for more care in shipments? The best and by far the fastest way is to send a photograph.

Our SX-70 Land camera makes it very simple. Just focus (as close as 10.4 inches without any attachments), aim and shoot. Minutes later you know you have a permanent, color record.

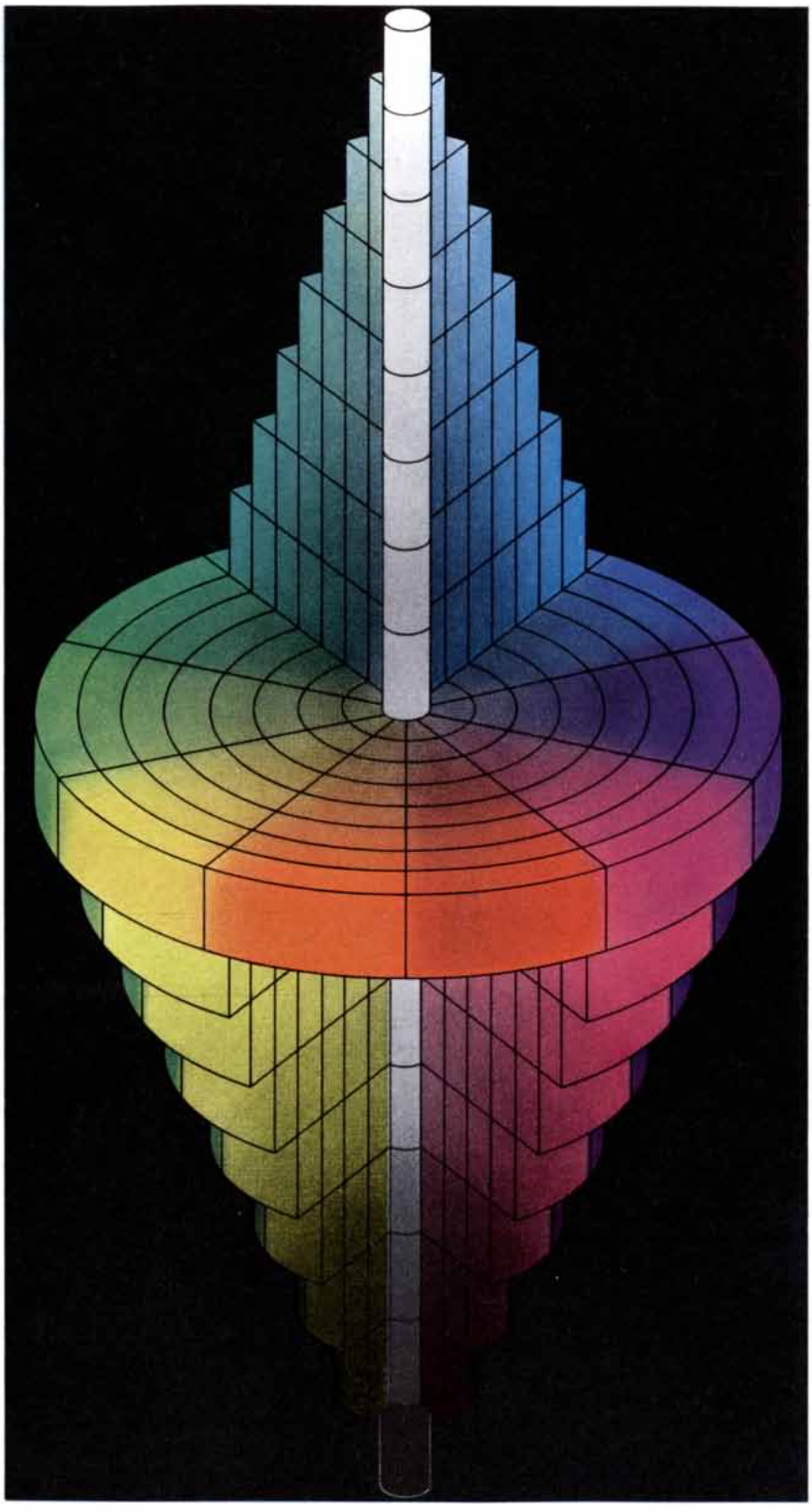
The SX-70 has become a new kind of transistorized business machine. Because it's the most compact, versatile, easy-to-use instant-picture camera made. Chances are it can make some function in your organization function better.

THE SX-70 CAMERA

In today's world, time is more valuable than ever. So the need for instant photography is greater than ever.

If you have a need for visual records, we can probably show you how to get them faster. Write to: Industrial Marketing, Dept. A305, Polaroid Corporation, Cambridge, Mass. 02139. Or you can call us collect: (617) 547-5176.

Polaroid Instant Photographic Systems



SURFACE COLORS VARY in lightness, saturation and hue. This variation is represented geometrically here in terms of a double cone in which lightness (white to black) varies along the axis, saturation along the radii and hue around the circumference. The highest degree of saturation is perceived for colors of medium lightness. The possible appearances of surface colors are not exhausted by the dimensions of lightness, saturation and hue. Film colors can also be described by a double-cone illustration, but with the attribute of brightness (dazzling to dim) in place of the attribute of lightness. A white or yellow film color will appear dimmer as the light intensity is reduced but will never look gray or brown. The attributes of brightness, saturation and hue exhaust the possible appearances of film colors.

intensity and wavelength composition of the incident illumination.

Subtle variations in the appearance of a color result from differences in the composition and texture of the surface. The same hue will not look the same on surfaces made of plastic, wood, metal, paper and cloth. In addition to the subtle color variations due to the material's composition and texture, new color attributes emerge with surface colors. Metallic colors such as gold, copper and silver are new color attributes that appear with the perception of a surface color. Surface colors may also vary along dimensions from glossy to matte, transparent to opaque and fluorescent to nonfluorescent.

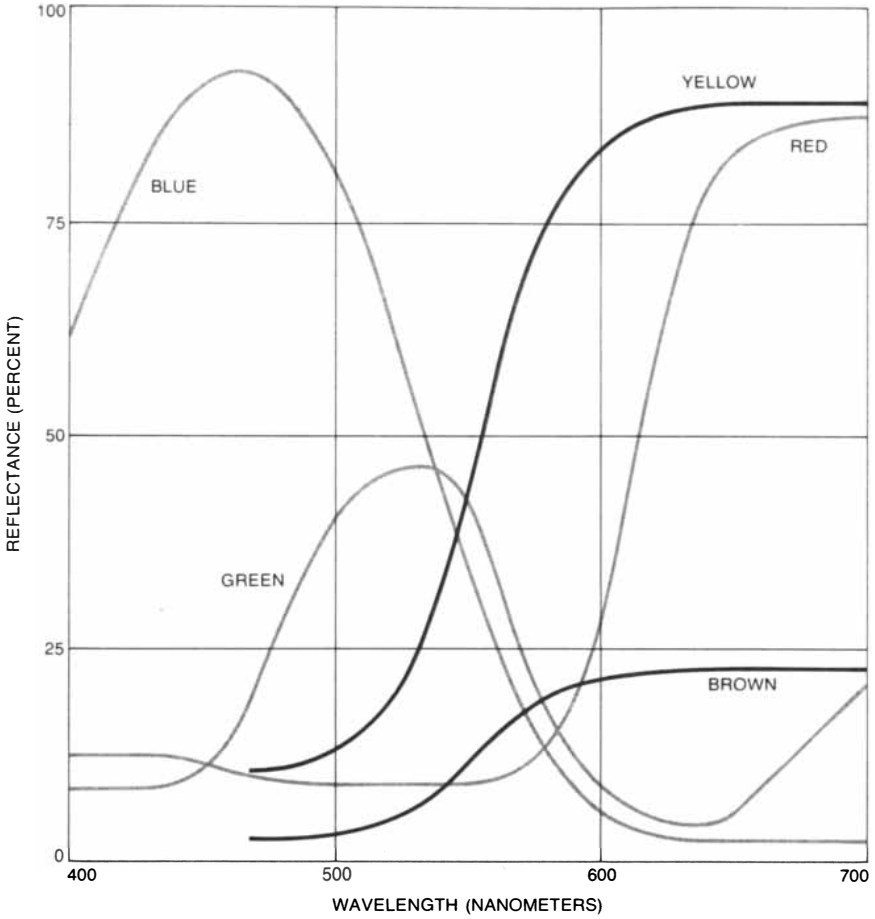
From a physical point of view these new color attributes are related to the complex ways surfaces reflect light. The degree of smoothness of a surface determines how the incident light is reflected; a perfectly smooth surface such as a mirror is said to reflect light specularly, whereas a rough surface reflects light diffusely. Most surfaces reflect light both specularly and diffusely in complex combinations. The perception of glossiness is correlated with the degree of specular reflection. The shiny appearance of certain objects is the result of highlights, or regions of strong specular reflection; the resulting glossiness of such a surface is called specular gloss. There are many other kinds of glossiness, which correspond to different patterns of specular and diffuse reflection. Luster and sheen, for example, are two further kinds of glossiness.

Characteristic surface appearances are also produced by differences between the light reflected from the outermost surface and the light reflected from the body of an object. An example of the distinction between surface and body color is the surface appearance called depth of finish. This attribute is associated with the absence of texture from the outermost surface of an object. The specular reflection of achromatic (colorless) light from the top surface of a polished piece of wood, for example, is also achromatic, but the diffuse reflection from the irregular wood fibers is colored, because the fibers absorb many of the wavelengths of the incident achromatic light. The human perceptual system separates the diffusely reflected colored light of the wood surface from the specularly reflected achromatic light of the layer of polish. One sees the wood surface below the shiny layer of polish. Metallic colors also result from specular reflection from the top surface of metals.

In metals such as gold and brass selective specular reflection results in highlights that are colored, unlike the colorless highlights of a glossy surface.

The perception of surface color involves the response of the visual system to patterns of stimulation. It is possible to determine the color quality of local areas of a metallic or glossy color, but these areas will not have the characteristic appearance associated with the entire surface. Local areas of a silver color appear as differing shades of gray; local areas of a gold color appear as differing hues of orange and yellow. The complex character of the perception of gloss is indicated by its dependence on perceiving the three-dimensionality of an object. Without the perception of an object such as a vase as being a curved surface the perception of gloss tends to disappear. Completion processes also play an important role in the perception of a glossy surface. Discrete highlights on a vase produce the perception of shininess over the entire vase. When the highlights are removed, the appearance of the entire surface changes and becomes matte [*see bottom illustration on next page*].

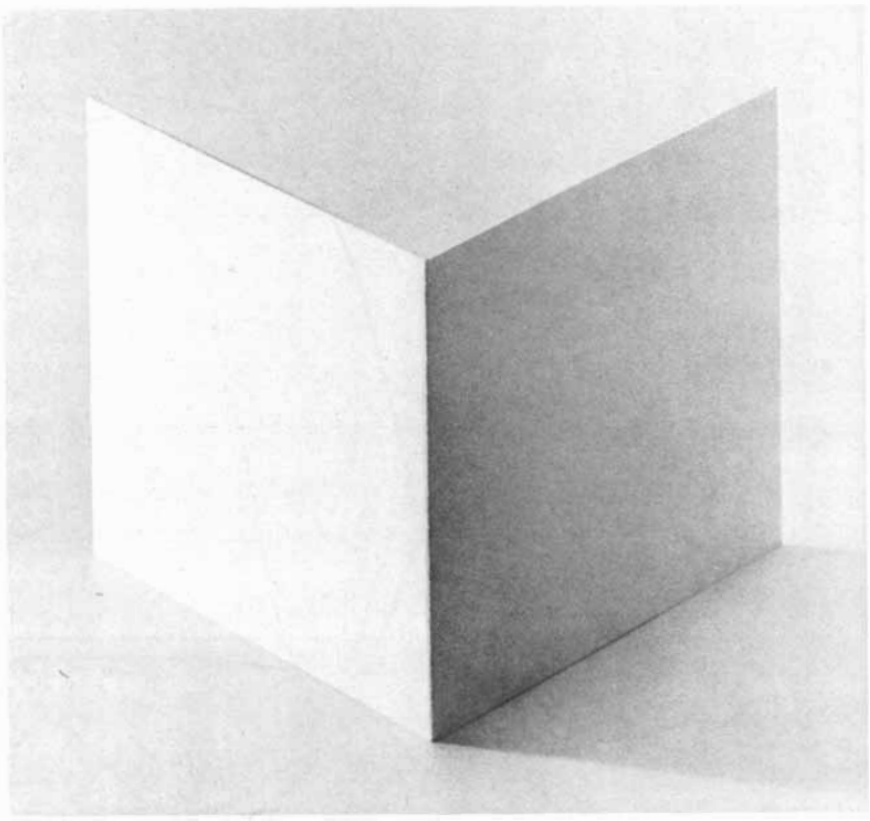
The general problem is to account for how a surface that reflects a distribution of luminances and hues yields a specific perception of color. I have suggested that the perception of a color as an attribute of a surface that reflects light nonuniformly can be analyzed as a two-stage process. First, sensory processes determine a pattern of neural signals in accordance with the distribution of luminances reflected. Second, as a result of the cue properties of stimuli these signals are organized by the visual system in terms of a single color with deviations seen as highlights, light spots and shadows. Owing to the shape, size, edge gradients and distribution of luminance variations, luminance differences varying from a particular value on a glossy surface are perceived as highlights and light spots rather than as differences in surface lightness. Thus a photograph of a shade-dappled object can be seen as one color, partly illuminated and partly in shadow. If one scans the entire image through a small aperture, the color of one part appears to be the color of the whole. The other parts of the object are seen as being too light or too dark. It is clear that one can take two attitudes in looking at such a picture. One can see the single color exhibited by the pattern of color variations on the object or one can see the individual colors making up the pattern.



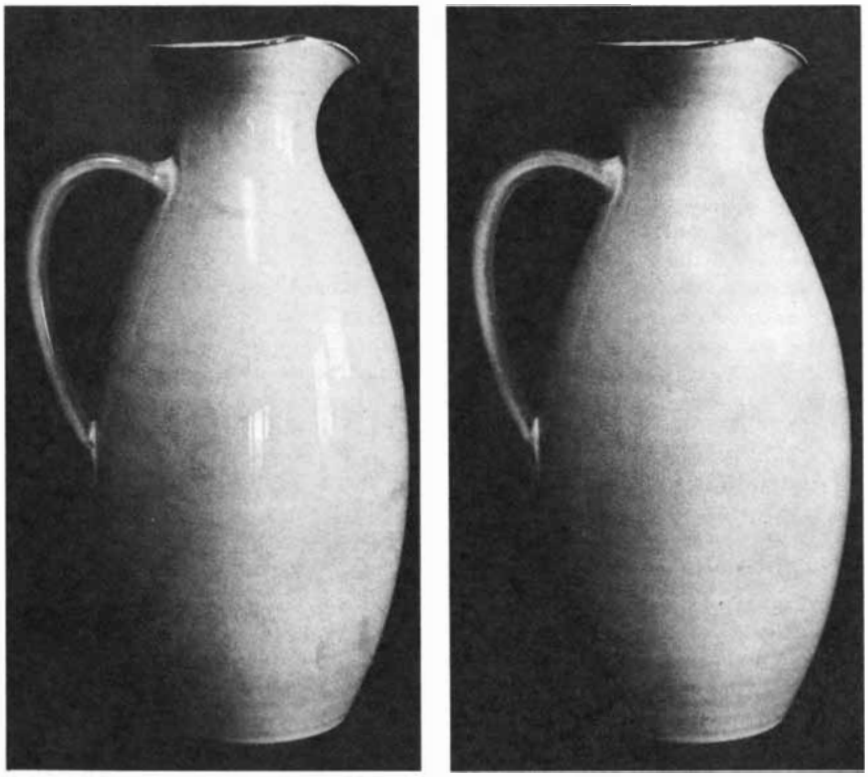
YELLOW AND BROWN SURFACES may reflect the same wavelengths of light and hence be identical in hue, yet they may be perceived differently; the apparent difference in this case results from the fact that the brown surface reflects less light than the yellow surface.

Surface colors typically involve complex perceptual structures. Consider a blue scarf thrown carelessly on a table. Looking at the scarf, one is able to identify the wool material out of which the scarf is made, the pattern of knitting that determines the texture of the scarf, the hue and lightness of the scarf, the overall illumination falling on the scarf, the distribution of light and shade, the folds and creases of the scarf determined by the way it has fallen. These aspects are tied together in an organized structure and are separable from the perception of the identity of the object as a scarf. At present it is not possible to specify in detail the processes by which the human visual system analyzes the information in the distribution of reflected luminances to produce a structured pattern. The processes appear to involve both specific cues such as the sharpness of contours [*see top illustration on page 69*] and organizational tendencies that lead the viewer to minimize lightness changes over a perceived surface [*see bottom illustration on page 69*].

An important characteristic of surface colors is their tendency to retain their daylight appearance in spite of changes in the intensity and wavelength composition of the incident illumination. The phenomenon of color constancy has been studied by many investigators, beginning in the 19th century with Hermann von Helmholtz and Ewald Hering. Helmholtz proposed that color constancy is the result of perceptual processes that take into account both the intensity and the hue of the incident illumination. According to Helmholtz, a weakly illuminated piece of gray paper is seen as gray and a strongly illuminated piece of black paper is seen as black even though the intensity of the light reflected by the black paper may be greater than that reflected by the gray paper. The reason is that the gray paper is judged to reflect a medium amount of light in proportion to the incident illumination, whereas the



FOLDED CARD illuminated from the left can be seen as uniformly white even though the part at right is in shadow, providing a simple demonstration of color constancy. Seen in false perspective (with fold away from camera) right side appears darker than the left side.

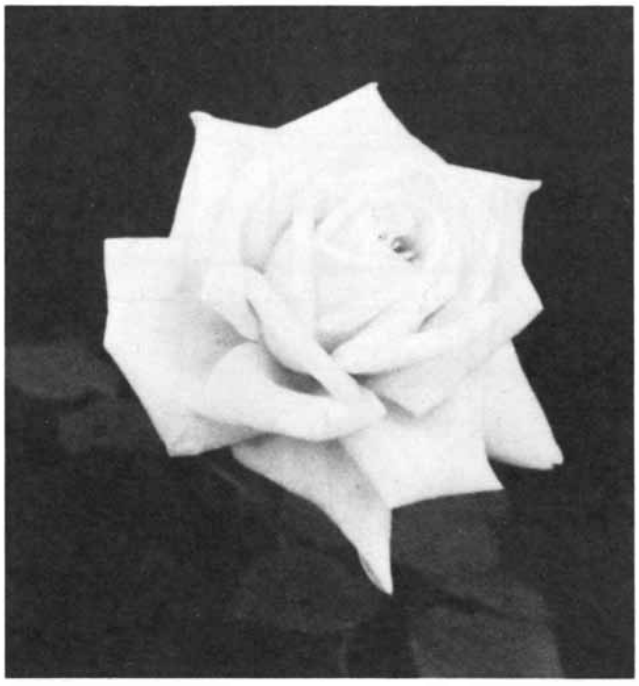


HIGHLIGHTS, or regions of strong specular reflection, on a three-dimensional object such as a vase help to make the entire surface of the vase appear glossy (*left*). When the photograph is retouched to remove the highlights, the surface of the vase appears matte (*right*).

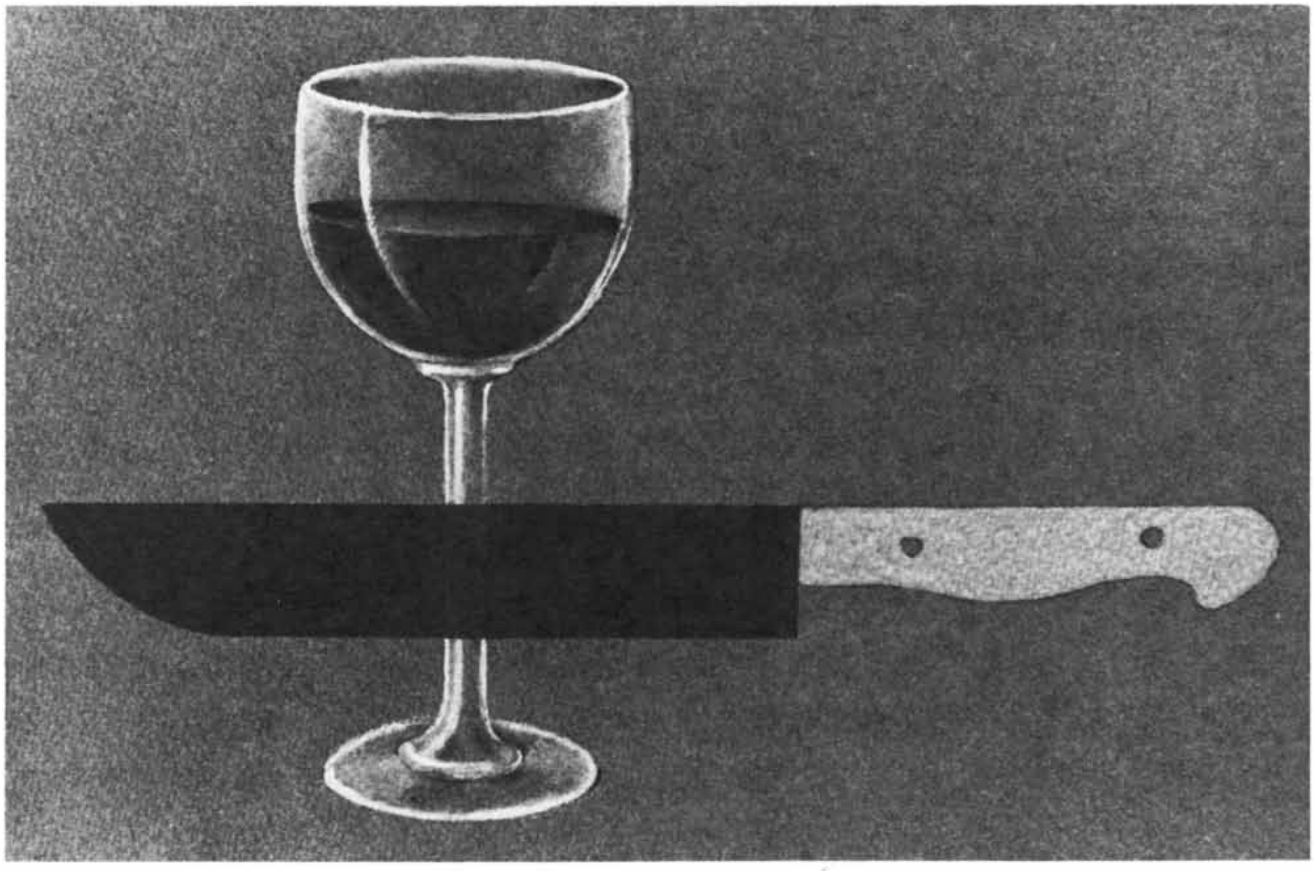
black paper is judged to reflect little light in proportion to the incident illumination [*see top illustration on page 72*].

Hering, on the other hand, argued that the major factors responsible for color constancy are the sensory processes of adaptation and contrast. He did allow secondarily for the effects of "color memory," which causes the color of a familiar object to be seen in accordance with its daylight color. According to Hering, a weakly illuminated gray is seen as gray and a strongly illuminated black is seen as black because contrast makes lightness dependent on the luminance ratio of an object to its background. A contrast ratio of background to object of 3 : 1 gives rise to gray; a contrast ratio of background to object of 18 : 1 gives rise to black [*see middle illustration on page 72*].

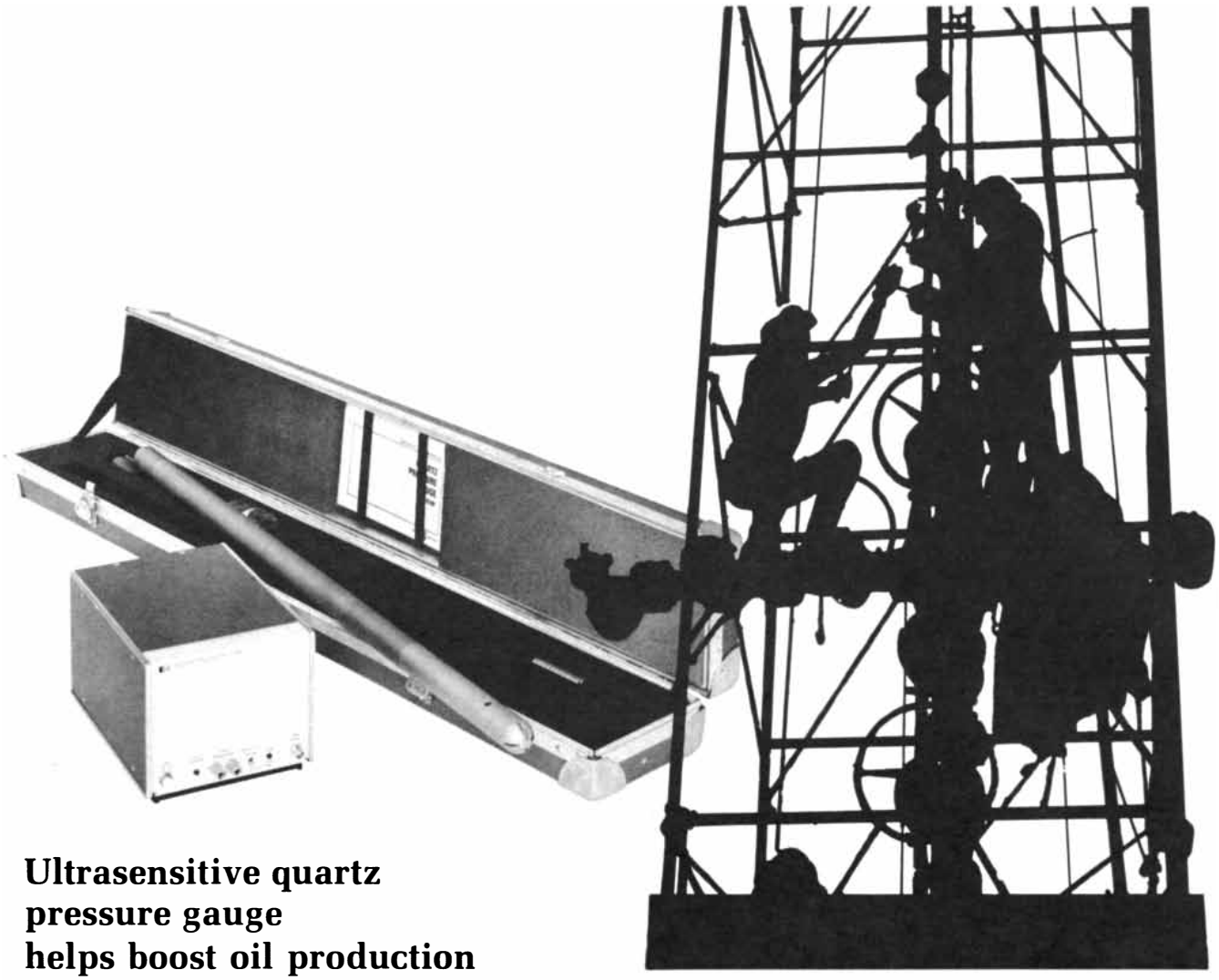
Both the sensory theory of Hering and the theory of Helmholtz that the perception of color depends on information concerning the illumination are able to account for many of the experiments on color constancy. Recent studies have clearly demonstrated, however, that cues for the illumination can affect color perception. In one such experiment I used a shadow to give the impression that a surface was strongly illuminated [*see bottom illustration on page 72*]. Two pieces of black matte cardboard were placed on a table in front of white backgrounds. Portions of the two black surfaces and their backgrounds were illuminated by projector beams. The two illuminated areas of the black surfaces served as targets. In the shadow condition the light fell partly on a corner of the surface and partly on the background above and beside the surface, producing a deep black shadow. In the no-shadow condition the light fell half on the black surface and half on the white background. People who judged the lightness of the two illuminated black areas judged the black surface in the shadow condition to be darker. The fact that a shadow gave rise to the perception of a darker color cannot be explained by the mechanism of lightness contrast. By that standard the shadow, because it was of lower luminance and lay between the black surface and the white background, should have decreased the contrast and caused the target casting a shadow to be seen as lighter than (or just as light as) the target without a shadow. In terms of the hypothesis that cues for the illumination can affect perceived lightness, the shadow provides direct visual information that the target is illuminated by a spot-



GRADUAL TRANSITION from light to dark characteristic of a shadow serves as one of the cues that enable the human visual system to perceive an object as a uniformly colored surface partly illuminated and partly in shadow (*left*). When this transition zone, or penumbra, is obliterated (*right*), the appearance of a shadow is lost and a gray spot is seen. The photograph reproduced here appeared originally in *Eye, Film and Camera in Color Photography*, by Ralph M. Evans, published by John Wiley & Sons in 1959.



CHANGES IN LIGHTNESS over a perceived surface tend to be minimized by the human visual system even when they run counter to experience. Viewed casually, the stem of the goblet in this sketch is usually regarded as being of constant lightness. Color constancy here requires that the viewer see the stem as standing in a reduced illumination behind a transparent knife blade. The sketch is copied from a 1968 article in *Acta Psychologica* titled "Perception, Past Experience and the Impossible Experiment," by G. Kanizsa.



Ultrasensitive quartz pressure gauge helps boost oil production

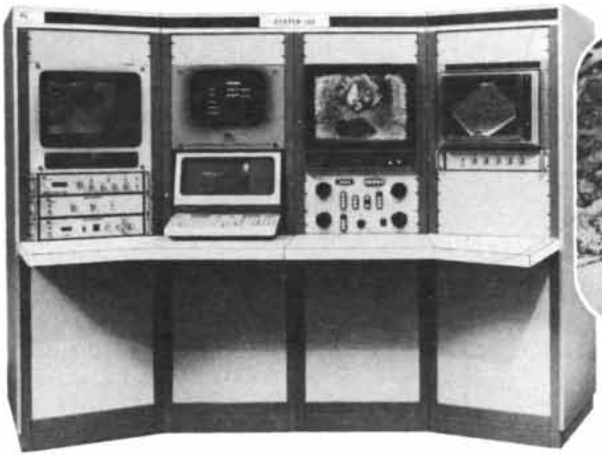
Able to measure a change as small as 0.01 psi at wellbore pressures up to 11,000 psi, the HP 2811B helps engineers evaluate reservoir parameters for optimal oil recovery.

One of the most effective methods of determining the production capacity of an oil reservoir is called pulse testing. By applying a series of pressure changes in a well and measuring the response in an adjacent observation well, engineers can determine the height and drainage characteristics of the underlying reservoir and thus calculate how to optimize its production.

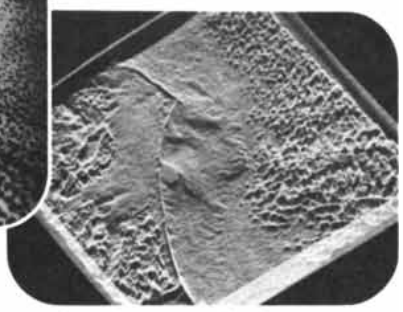
Although pulse testing theory has been thoroughly understood for ten years, it has not been widely used, for want of an adequately sensitive pressure gauge. With ordinary instruments, large pulses have to be applied over long cycle times in order to detect the greatly attenuated response in the observation well. The test therefore takes a long time, typically one or more weeks, and the loss in oil production is considerable since both wells are "shut-in" during the test.

The HP 2811B Quartz Pressure Gauge has changed things dramatically for the better. Because it can accurately measure a change as small as 0.01 psi at any reservoir pressure up to 11,000 psi—compared to 0.5 psi with ordinary instruments—the pressure changes can be much smaller, cycle times much shorter, and the duration of the test cut to a few days, depending on reservoir conditions.

Reservoir engineers find many other reasons to prefer the 2811B as a "bottom hole" pressure gauge for pulse testing. Strip chart recorder and digital printer options provide an instant and direct readout of pressure changes on *the surface*, even when the measuring probe lies 20,000 feet down the hole. The gauge maintains resolution and accuracy at any well pressure to 12,000 psia and any temperature to 300°F. And it holds its calibration for at least a full year despite mechanical vibration and rough handling. Cost of the complete gauge without recording options is $11,375*.



Electronic enhancement of Pamlico Sound, North Carolina, from a satellite photograph.



A three-dimensional enhancement of the same view.

Digital image processing system unlocks earth's secrets from satellite data.

Stanford Technology Corporation chooses HP 3000CX computer system to convert radio signals from the LANDSAT (ERTS) satellites into useful pictures of the earth's resources.

One of NASA's least known spacecraft is well on the way to making some of the most important "civilian" contributions of the space program.

Since 1972, the first Earth Resources Technology Satellite has been looking at the whole earth through a multispectral optical scanner. A second LANDSAT satellite has been in service since January 1975. When properly processed, the digital data transmitted to earth from these satellites provides spectral "signatures" of classes of objects on earth that can be used to inventory many of the world's resources. Agronomists have used the pictures to measure the total acreage of various crops and to project their yields; foresters, to detect timberland insect infestation; planners, to outline land-use patterns and flood-prone areas; geologists, to locate mineral deposits.

The potential usefulness of these pictures in many fields has created a great demand for equipment to interpret them. A new development by Stanford Technology Corporation—the System 101—is being offered to meet this demand.

This powerful new multi-user digital image processing system is configured around the HP 3000CX, chosen by STC engineers as the computer system best suited to the application.

While the HP 3000CX is fast and powerful enough to satisfy the full range of LANDSAT requirements, it is much easier to use than larger computers, especially by nonspecialists. Many scientists can use it at the same time, some processing images on-line while others develop programs. An extremely simple language with a "menu-prompting" mode is provided for inexperienced users, while advanced users can use a high-level language with efficient command lists.

The STC System 101 should go a long way in reducing the LANDSAT image-processing bottleneck. For more information on System 101, write or call Stanford Technology Corporation in Mountain View, California 94043.

System prices for the HP 3000CX start at $99,500*.

For more information on these products, write to us, Hewlett-Packard, 1503 Page Mill Road, Palo Alto, California 94304.

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries

Mail to: Hewlett-Packard, 1503 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP 3000CX Computer System
( ) HP 2811B Quartz Pressure Gauge

Name _____ Title _____
Company _____
Address _____
City _____ State _____ Zip _____

*Domestic USA prices only.

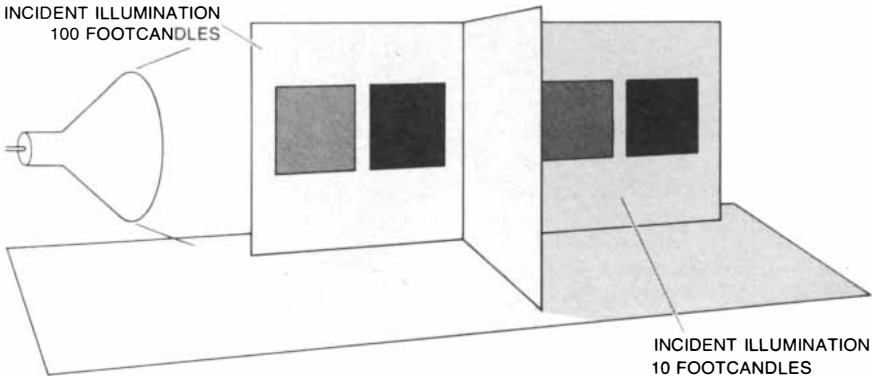
00547

light and gives rise to the perception of a darker surface in a strong illumination rather than of a lighter gray surface.

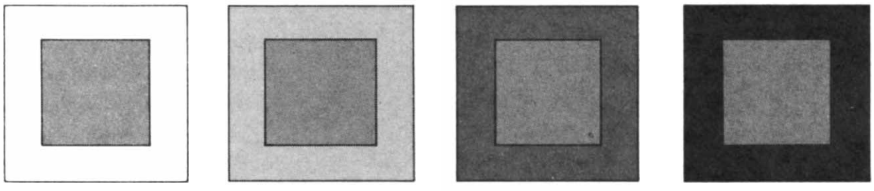
Experiments showing that the perceived lightness of a surface is affected by the apparent spatial position of a surface further support the hypothesis that information about illumination can affect perceived lightness. A trapezoid placed upright on a table can be seen as either a trapezoid standing upright, when cues for its slant are present, or a rectangle lying flat on the table, when cues for the slant of the trapezoid are eliminated [*see illustration on opposite page*]. Julian E. Hochberg and I found that a trapezoid physically upright was perceived as darker when it was made to appear at right angles to the direction of illumination than when it was made to appear parallel to the direction of the illumination. For example, when the illumination came from overhead, the trapezoid appeared darker when it was seen as a rectangle lying flat than when it was seen as a trapezoid standing upright. When the trapezoid was seen as an upright trapezoid, parallel to the perceived direction of illumination, the trapezoid appeared as a shadowed surface. When it was seen as a flat rectangle perpendicular to the perceived direction of illumination, the trapezoid was no longer seen as shadowed but appeared as a darker surface color.

Another instance of the dependence of perceived lightness on apparent position is illustrated by the example of the folded card given near the beginning of this article. Experiments indicate that when the folded card is seen as it physically is (that is, with the fold toward the observer), observers see the two sides of the card as having the same lightness, the left side illuminated and the right side in shadow. When the folded card is seen in false perspective (that is, with the fold away from the observer), observers see the two sides of the card as differing in lightness, with the right side darker than the left. The results of experiments on the effects of the apparent position of a surface clearly show that when a luminance variation is no longer consistent with the perception of a shadow, the "percept" changes so that the area of lower luminance is perceived as an area of lower surface reflectance (that is, a darker surface color).

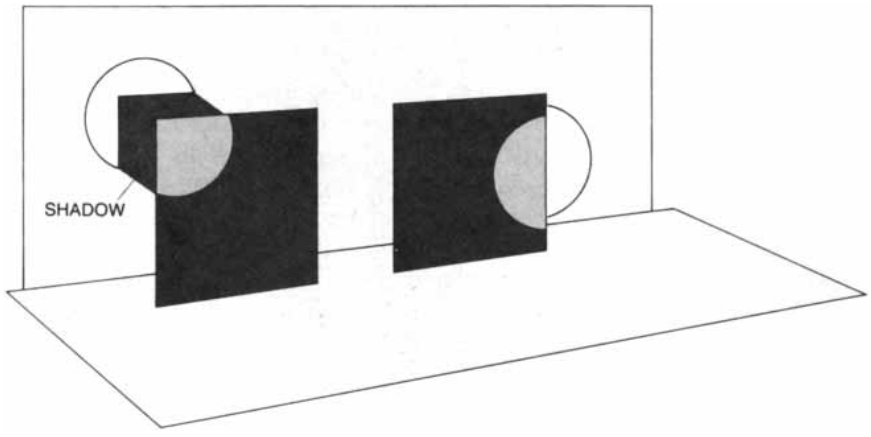
An interesting example of color constancy involves transparent colors, which depend on the perception of one surface behind another. Look at a white surface through a dark film (such as a neutral-density filter, which transmits only a portion of the light striking it) and the surface will be seen as a white in a reduced illumination rather than as a darker gray. The area of intersection of each pair of colored circles on the left side of the illustration on page 75 is the result of combining paints of the two colors. In each pair the area of intersection can be seen either as the mixture color when both colors are seen in the same plane or as the color of one circle seen through the transparent color of the other circle when the circles are seen in



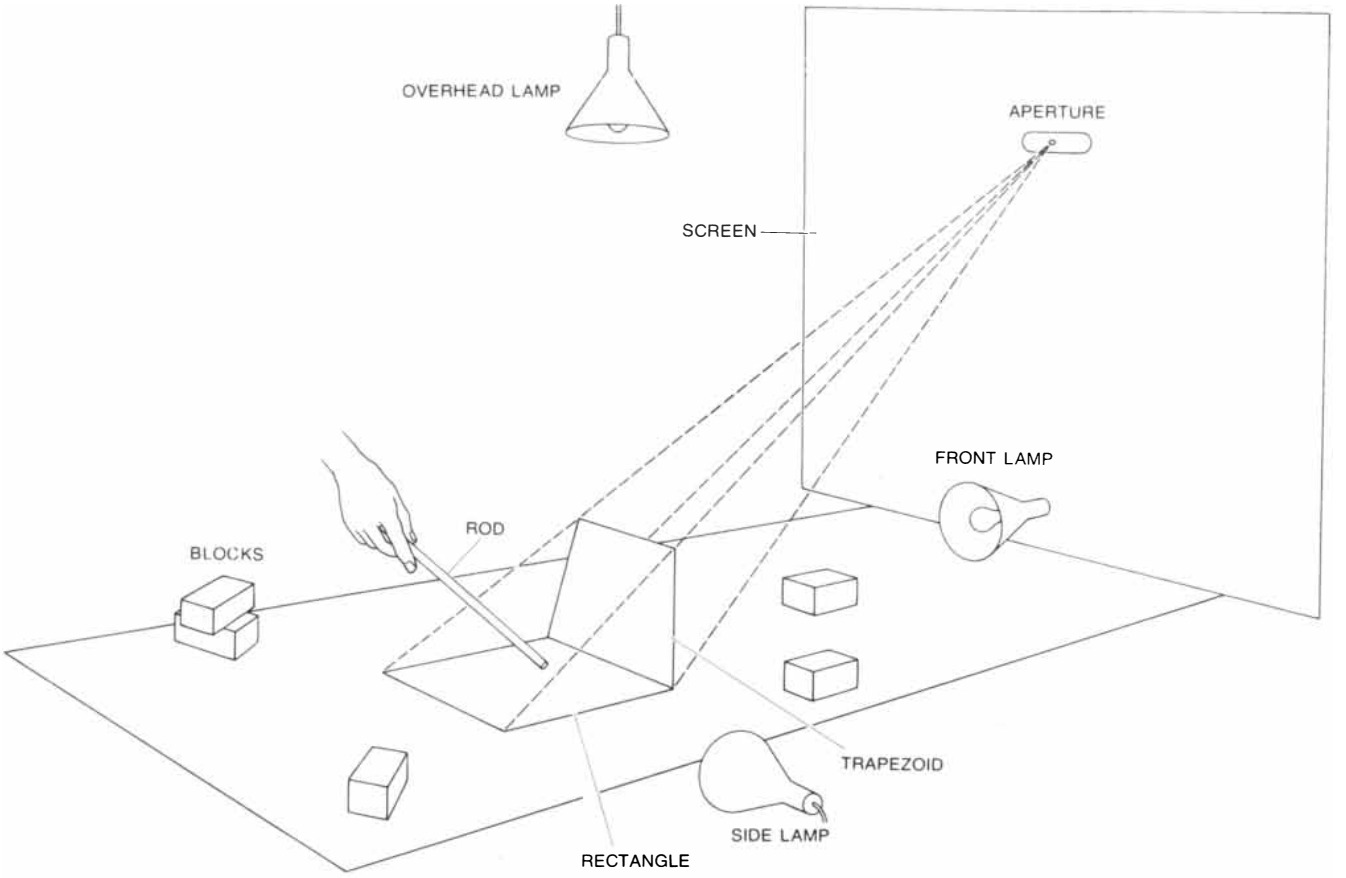
COLOR CONSTANCY leads one to perceive the white, gray and black surfaces at left as approximately the same as the white, gray and black surfaces at right, even though the intensity of the light reflected from the two sides of the panel differs drastically. (In an actual test the light intensities reflected from the white, gray and black surfaces on the illuminated side were 80, 30 and five footlamberts; the light intensities reflected from the identical surfaces on the shaded side were eight, three and .5 footlamberts.) Two different explanations of this phenomenon were offered in 19th century by Hermann von Helmholtz and Ewald Hering.



CONTRAST EFFECT was believed by Hering to be at the root of the phenomenon of color constancy. In this illustration the small squares are all the same medium gray and reflect identical light intensities. Their perceived lightness differs because they are viewed against backgrounds of different luminance. According to Hering, a contrast ratio between background and object of 3 : 1 gives rise to gray; a contrast ratio of 18 : 1 gives rise to black.



RECENT EXPERIMENT by the author at the University of Oregon showed that a shadow cast on a background can give a viewer the impression that the surface casting the shadow is strongly illuminated by a spotlight, causing the surface to appear darker than it otherwise would. The fact that the shadow condition (*left*) gave rise to perception of a darker color than the no-shadow condition (*right*) cannot be explained in terms of lightness contrast.



APPARENT POSITION OF A SURFACE can affect the perceived lightness of the surface, as demonstrated in this experimental arrangement by Julian E. Hochberg and the author. People who looked with one eye through the small aperture in the screen tended to see the upright trapezoid as a rectangle lying flat on the table. When a rod was waved behind the trapezoid, or when the scene was viewed with both eyes through a larger opening in the screen, the trapezoid was seen correctly as standing upright on the table. When the distribution of shadows indicated that illumination came from overhead, the target surface appeared darker when seen as a flat rectangle than when seen as an upright trapezoid. When the distribution of shadows indicated that illumination came from the front, the target surface appeared darker when seen as an upright trapezoid than when seen as a flat rectangle. When the distribution of shadows indicated that illumination came from the side, the target surface did not change in lightness when seen in either position.

different planes. For example, one can see the area of intersection of the yellow and blue circles as a green, as a yellow circle behind a transparent blue circle or as a blue circle behind a transparent yellow circle. Although it is more difficult, one can see the area of intersection of the blue, yellow and magenta circles as black or as blue, yellow and magenta circles overlapping one another.

The foregoing suggests that transparent colors are perceived in terms of subtractive rather than additive color mixture. Additive color mixture is involved when by mixing lights one produces the hues that result from the simultaneous excitation of the three different types of cones (color receptors) in the retina. For example, the addition of yellow and blue hues produces an achromatic color: white or gray. Subtractive color mixture predominates when paints of different colors are mixed. When yellow and blue paints are mixed, subtractive color mixture produces a green paint. The yellow paint particles strongly absorb the short blue-producing wavelengths; the blue paint particles strongly absorb the longer yellow-producing and red-producing wavelengths. Thus the middle green-producing wavelengths are reflected most strongly by the mixture of yellow and blue paint particles and one sees green. The suggestion that the perception of color transparency follows the rules of subtractive rather than additive color mixture means that in the case of the colored circles on the right side of the illustration on page 75 it is not possible to see the yellow area of intersection of the orange and green circles as an orange in front of a green, or the white area as overlapping orange, green and blue-purple circles.

The subtractive nature of the perception of color transparency suggests that transparency effects have their bases in everyday experience. For example, subtractive color mixture is involved when a person looks at a surface through a pane of colored glass. When one looks at a blue surface through a pane of yellow glass, the light stimulating the eyes is whatever remains after the selective absorption by the blue surface and the yellow glass. Hence the eyes are stimulated predominantly by green wavelengths. The perception of transparency is favored by stimuli that give rise to the perception of two surfaces separated in depth; the phenomenon is also favored by the tendency of the human perceptual system to see complete and closed figures. The hypothesis that transparency is perceived in accordance with subtractive color mixture implies that transparency effects are not based on splitting the color signals that reach the brain into the component hue sensations signaled by

It's not new. It's proved.

Catalytic converter, standard equipment on 1975 model GM cars.

After more than a billion miles on the road, GM's catalytic converter has become a world standard in fuel-saving pollution control devices. Auto-makers in Europe, Asia and North America have contracted to buy converters from GM.

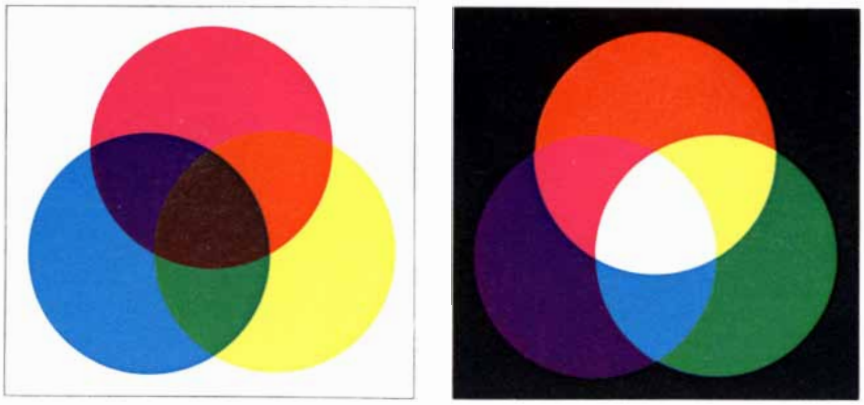
In operation on 1975 cars, the converter has proved that exhaust emission of carbon monoxide and hydrocarbons can be reduced by about 50% from the already lowered 1974 levels. And at the same time gas mileage on GM cars has been increased in city driving by 28% on a sales-weighted average, according to EPA calculations.

The converter is the kind of startlingly effective advance in automobile engineering that challenges other technology. It demands lead-free, low sulfur fuel. It leaves behind a lot of other pollution control concepts that don't have its qualities of durability and its fuel-saving characteristics. So the converter has come in for a bit of what we call "wishful criticism."

The fuel-saving, ecologically sound catalytic converter. Standard equipment on 1975 cars from General Motors, a world leader in automotive pollution control technology.

General Motors

Chevrolet, Pontiac, Oldsmobile, Buick, Cadillac, GMC Truck



TRANSPARENT COLORS, which depend on the perception of one surface behind another, are seen in accordance with subtractive rather than additive color mixture, indicating that transparency effects are based on everyday experience. Subtractive color mixture predominates when one combines paints or filters of different colors (*left*). Superimposing blue, yellow and magenta paints on a white field in pairs produces orange, green and blue-purple. Where all three paints overlap, black is seen. Additive color mixture results from combining lights of different colors (*right*). Superimposing orange, green and blue-purple beams of light on a white screen in an otherwise dark room in pairs produces blue, yellow and magenta. Where all three beams overlap, they produce white. Color printing with inks involves both subtractive and additive color mixture and can only approximate the two conditions.

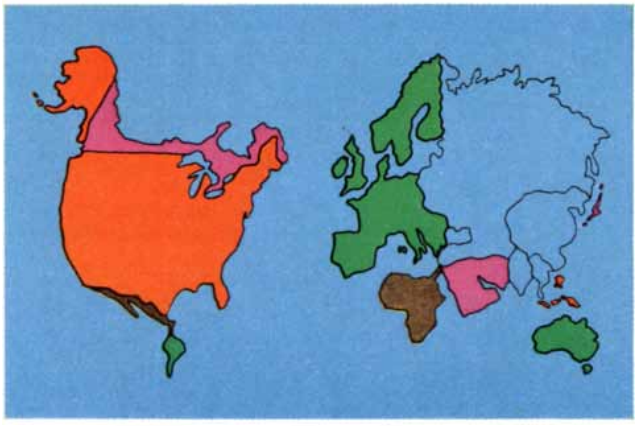
the retinal color receptors. Rather, they arise from specific experiences with colors as they are encountered in everyday settings. An alternative hypothesis—that transparency effects are based on additive color mixture—is supported by some investigators [see "The Perception of Transparency," by Fabio Metelli; SCIENTIFIC AMERICAN, April, 1974].

How do cues for illumination affect the perception of color? To simplify the discussion I shall consider only how information about the intensity of the illumination affects the perception of lightness. One explanation appeals to a kind of unconscious computation in which the visual system combines information about the intensity of the light reflected from a surface and the intensity of the illumination in perceiving lightness. This hypothesis asserts that processes underlying the perception of lightness produce the same results as those given by the "albedo equation," according to which the albedo, or perceived lightness, is equal to the ratio of the intensity of the reflected light to the registered illumination. This explanation is now found to be unsatisfactory. First, the perception of lightness fails to vary as a linear function of the perceived illumination, as is implied by the albedo equation. For instance, the results of experiments with the folded card and the trapezoid-rectangle figures show that the perceptions of illumination and lightness do not covary in the manner required by the albedo equation. Second, information about surface illumination does not in all instances affect the perception of lightness. The results of the recent experiments suggest that it is stimulus information that creates the impression of a "special" illumination (a decreased illumination due to a shadow or an enhanced illumination due to a spotlight) affecting the perception of lightness.

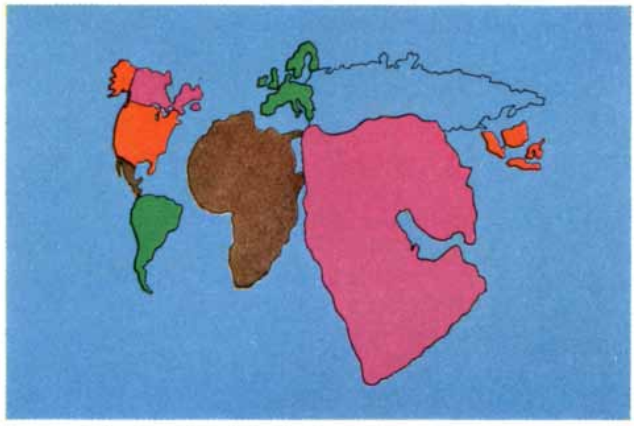
The results of these experiments suggest a somewhat looser relation between the perceptions of illumination and lightness. If surfaces are perceived as being uniformly illuminated, and the effects of color memory are ruled out, variations of luminance are perceived as differences in surface lightness, not in illumination; this enables lightness perception to be expressed as a function of sensory processes [see "The Perception of Neutral Colors," by Hans Wallach; SCIENTIFIC AMERICAN, January, 1963]. If surfaces are perceived as being nonuniformly illuminated, however, illumination cues, object cues, organizational tendencies and a viewer's attitude can determine whether a person will see an area of altered surface luminance as an area of altered illumination or as a difference in surface lightness. A revised theory of the perception of surface color, in short, must go beyond the albedo hypothesis to take into account the fact that cues for the illumination represent only one of a number of factors that can affect how one perceives the degree of lightness of a colored surface.

America's coal: a gold

The U.S. has more energy in coal than in oil and gas combined. Here's what Exxon is doing to help put coal's energy to work today.



The world of coal. The map above shows what the world would look like if the size of each country were in proportion to the coal reserves under its land. America—with nearly half the free world's coal—dwarfs Europe and the Middle East. Experts estimate that at the rate the United States uses coal today, these reserves could help keep us in energy for the next two hundred years. This could help supplement dwindling supplies of U.S. gas and oil—and more importantly, provide our country with an energy source that's right here at home.



The world of oil. This map shows where the world's oil is located. Most of it is in the Middle East. By comparison, however, the energy in U.S. coal is twice the amount of energy in Middle East oil. Exxon anticipated the growing demand for coal several years ago. That's why Exxon is continuing its program aimed at helping America take greater advantage of this vast energy resource. Exxon scientists and engineers are working to find better ways to mine coal, cleaner ways to burn it, and new ways to put coal to work.

An example of land reclamation. Exxon is firmly committed to reclaiming land which may be disturbed by our mining. We have no operational coal mines in the West right now. But, our uranium mine near Casper, Wyoming, is an excellent example of how our reclamation program works. Since operations began, Exxon has been continually contouring, planting and reseeding the disturbed areas with yellow clover, crested and western wheat grasses. Right now, it's difficult to tell the natural landscape from the reseeded landscape. When we've finished mining here, the land will be at least as productive as it was before we came.

mine of energy.

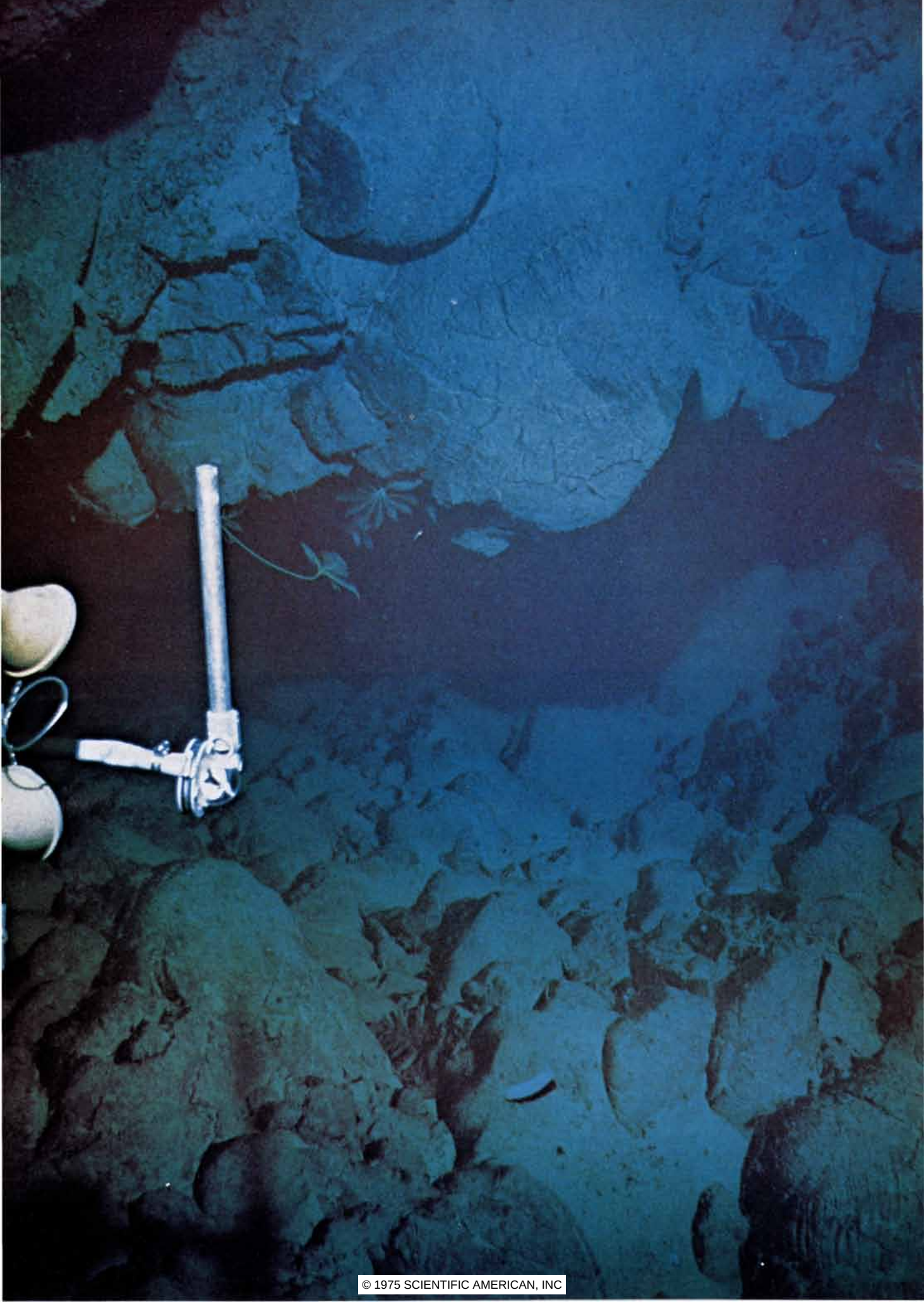
A new kind of coal town. To most visitors, the Court House at Carlinville, Illinois, typifies the picturesque beauty of this Midwest community. In 1970, Exxon opened a new mine near Carlinville. Today, the town is still as clean and picturesque as ever. And a lot more prosperous. Employment is up 22 percent, with over 300 new and high-paying jobs. Recently, leaders from other prospective coal towns visited Carlinville to see how coal mining and a good city can thrive side by side.

America's biggest coal users. Most coal used in America today is burned by electric power plants like the one above. These plants consumed about 400 million tons of coal last year. By 1985, this figure could jump to nearly 700 million tons. This surge in coal consumption will go a long way toward freeing gas and oil for other purposes. And it will also mean thousands of new jobs for people here in the United States.

New mines, new miners, new methods. Putting a modern coal mine into production calls for a lot more than a pick and shovel. Many years must be spent in planning and development. New mining techniques must be tested. New mining talent must be trained. New transportation systems must be set up. All so that when a new Exxon mine opens it will produce coal in the most efficient and environmentally responsible way.

Cleaner energy from coal. Exxon is deeply involved in pollution control research. The TIGER van—or Traveling Industrial Gaseous Emission Research vehicle—is an important part of a program Exxon is conducting for a U.S. government agency. The TIGER van's job is to monitor and record emissions from coal-fired boilers and power plant stacks. The data collected is used by Exxon engineers to test new coal combustion techniques which may help bring America more energy without more pollution.

EXXON



The Floor of the Mid-Atlantic Rift

Last summer U.S. and French submersibles explored a rugged area at some 8,400 feet where lava wells up and the ocean floor moves outward, bearing with it the continents to the east and the west

by J. R. Heirtzler and W. B. Bryan

Some 20 years ago it became apparent that a continuous range of undersea mountains twists and branches for a total length of some 40,000 miles through all the world's oceans. This system of mid-ocean ridges is the site of frequent earthquakes with shallow foci. Geologists had no satisfactory explanation for these earthquakes, and they were even more mystified by the fact that samples of the ocean floor were geologically young. No samples seemed to be more than about 135 million years old, and the closer to the ridge the samples were taken, the younger they were. These findings, together with other discoveries (such as the regular alternation of magnetic polarity in broad bands of the ocean floor running parallel to the ridge), finally led to the hypothesis that the mid-ocean ridge is actually a system of parallel ridges centered on a continuous rift in the ocean floor, and that the floor itself is everywhere spreading outward from the rift. As the rift widens it is filled with lava that wells up from the earth's mantle. The ridges mark the edges of huge plates in the earth's crust that bear the continents. As the sea floor spreads, the continents at the margins of the oceans are either borne apart or compressed into new configurations.

Thus the mid-ocean-rift system reflects processes of the most fundamental importance in the evolution of the earth. Geologists now believe that the processes responsible for the rift system as we see it today have been creating and modifying the crustal plates throughout much of the four to five billion years of the earth's history. Although many details remain to be explained, the general concept of the creation of crust at the mid-ocean ridges has become widely accepted in less than a dozen years and forms the basis for much of present-day geological reasoning.

As has been the case with the spectacular growth of knowledge about the moon and the other planets, the expansion of our knowledge of ocean-bottom geology has been closely connected with technological advances that have made possible the increasingly rapid collection of data. Although much has been learned about the mid-ocean ridges by remote-sensing techniques, by drilling and by random sampling from surface ships, it finally became obvious that direct manned observation, by means of special submersible vessels, would ultimately be essential for any real understanding.

Starting late in 1971 a program was initiated by several interested groups to develop and harness specialized instruments and techniques for a detailed study of the central part of the Mid-Atlantic Ridge. A region of the ridge some 400 miles southwest of the Azores was selected for examination both because it is an area where one can expect good weather and because the port of Ponta Delgada in the Azores offered a convenient base of operations. The U.S. and France took leading roles in the effort because each had had experience with research submersibles of the type that would be needed for conducting manned observations in the later phases of the program. The cooperative venture was dubbed FAMOUS, for French-American Mid-Ocean Undersea Study.

Although the basic technology for deep-sea manned exploration was well developed, further refinements were needed. The French submersible *Archimède* had already gone much deeper than the 8,000-to-10,000-foot depths that would be encountered in the selected study area, but the *Archimède* was bulky, difficult to maneuver and provided only limited visibility. The U.S. submersible *Alvin* had already scored several technical successes in the Atlantic, including the recovery of the hydrogen bomb lost off the coast of Spain. Although the *Alvin* was small, maneuverable and afforded excellent visibility through its three portholes, it was limited to depths of less than 6,000 feet, too shallow for the depths of the Mid-Atlantic Ridge. Accordingly the French undertook to build a new submersible, the *Cyana,* similar in size and capability to the *Alvin,* and the U.S. fitted the *Alvin* with a new titanium pressure hull that would allow dives to at least 10,000 feet.

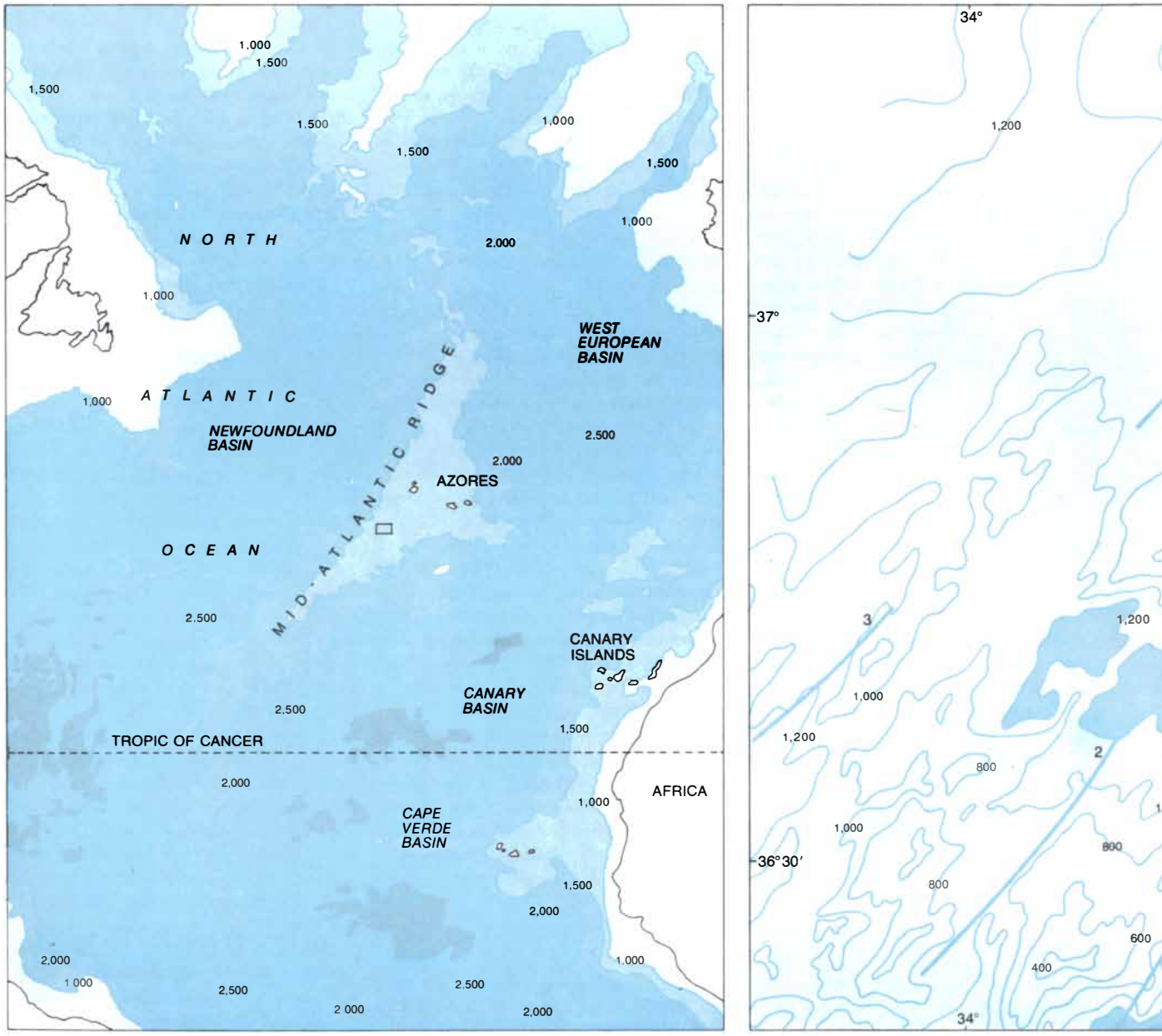
In further preparation for the venture plans were made to close an "observation gap" that would otherwise exist between the scale of detail that would be recorded from the submersibles, a scale measured in centimeters and meters, and the scale of observations that were then possible from surface vessels, a scale measured in hundreds of meters and kilometers. Strategies for closing the observation gap included echo sounding

TENSION FRACTURE in the floor of the Mid-Atlantic Rift was photographed with an automatic camera mounted on the research submersible *Alvin*. Along the sides of the fissure are numerous "pillow lavas," which are formed when lava is extruded into water and rapidly cooled. At the left is part of the *Alvin*'s external gear; the round object at the bend of the pipe is a compass. Flowerlike objects at top of the fissure are crinoids, invertebrate animals attached to the bottom. The *Alvin* was one of three submersibles that explored the rift as part of Project FAMOUS (French-American Mid-Ocean Undersea Study).

with a high-resolution narrow acoustic beam, photography with new techniques and the use of a variety of deep-towed geophysical instruments. All these methods would require the ability to position ships to an accuracy of a few tens of meters rather than the one kilometer that is usual in deep-sea navigation.

By 1971 there was no longer any doubt that our investigation would be confronted with detailed problems of volcanic geology. It was hard to realize that barely a dozen years earlier serious students of abyssal geology could still speculate that elevated portions of the mid-ocean-ridge system, such as the East Pacific Rise and the Azores Plateau, were remnants of sunken continental land areas. The deep central valley in the Mid-Atlantic Ridge and its transverse fracture zones, so well known today, were just beginning to be defined. The presence of outcrops of volcanic rock on the Mid-Atlantic Ridge was firmly established in 1961 by Earl Hayes and his co-workers on cruise No. 21 of the Woods Hole Oceanographic Institution vessel *Chain,* when they succeeded in photographing and sampling fresh, glassy volcanic rock at 28 degrees 53 minutes north latitude, due west of the Canary Islands. Over the next few years similar studies made by ships from various oceanographic institutions confirmed the presence of essentially similar volcanic rock on the crests of mid-ocean ridges in the South Atlantic, the Pacific and the Indian Ocean. Bizarre as the idea seemed at first, it was becoming evident that the mid-ocean-ridge system was nothing less than a vast unhealed volcanic wound.

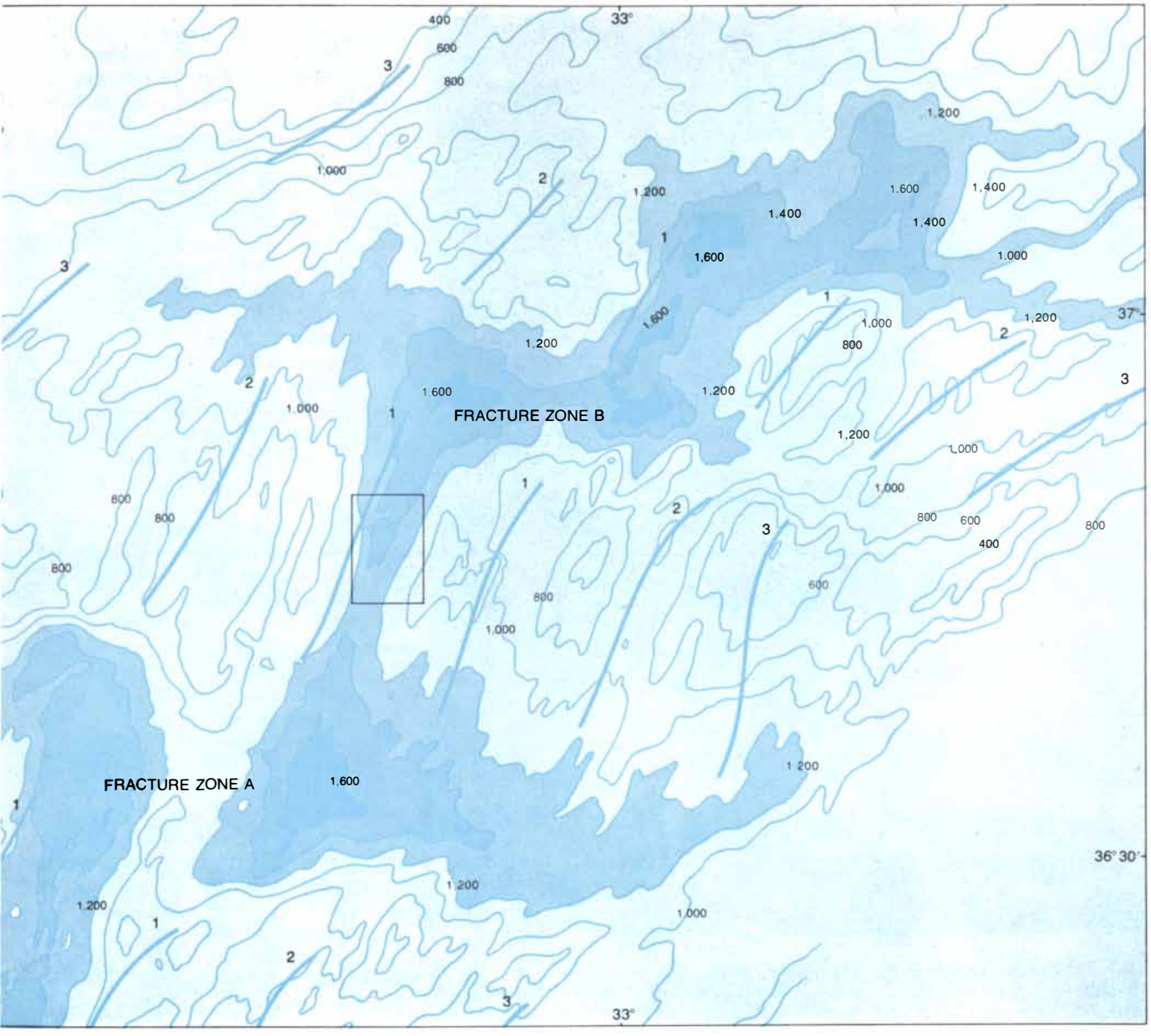
Even more remarkable, it was learned that finely divided particles of iron oxide



MID-ATLANTIC RIDGE (*map at left*) **is part of a continuous system of mid-ocean ridges some 40,000 miles long. Lava extrusions are centered on fissures along the crest of the ridges. These linear zones of extrusion mark the boundaries of huge lithospheric plates that are slowly moving apart, carrying the continents with them. As the lava solidifies, particles of iron oxide in the molten rock become aligned with the earth's magnetic field. Because of periodic reversals in the polarity of the earth's field the solidified and mag-**

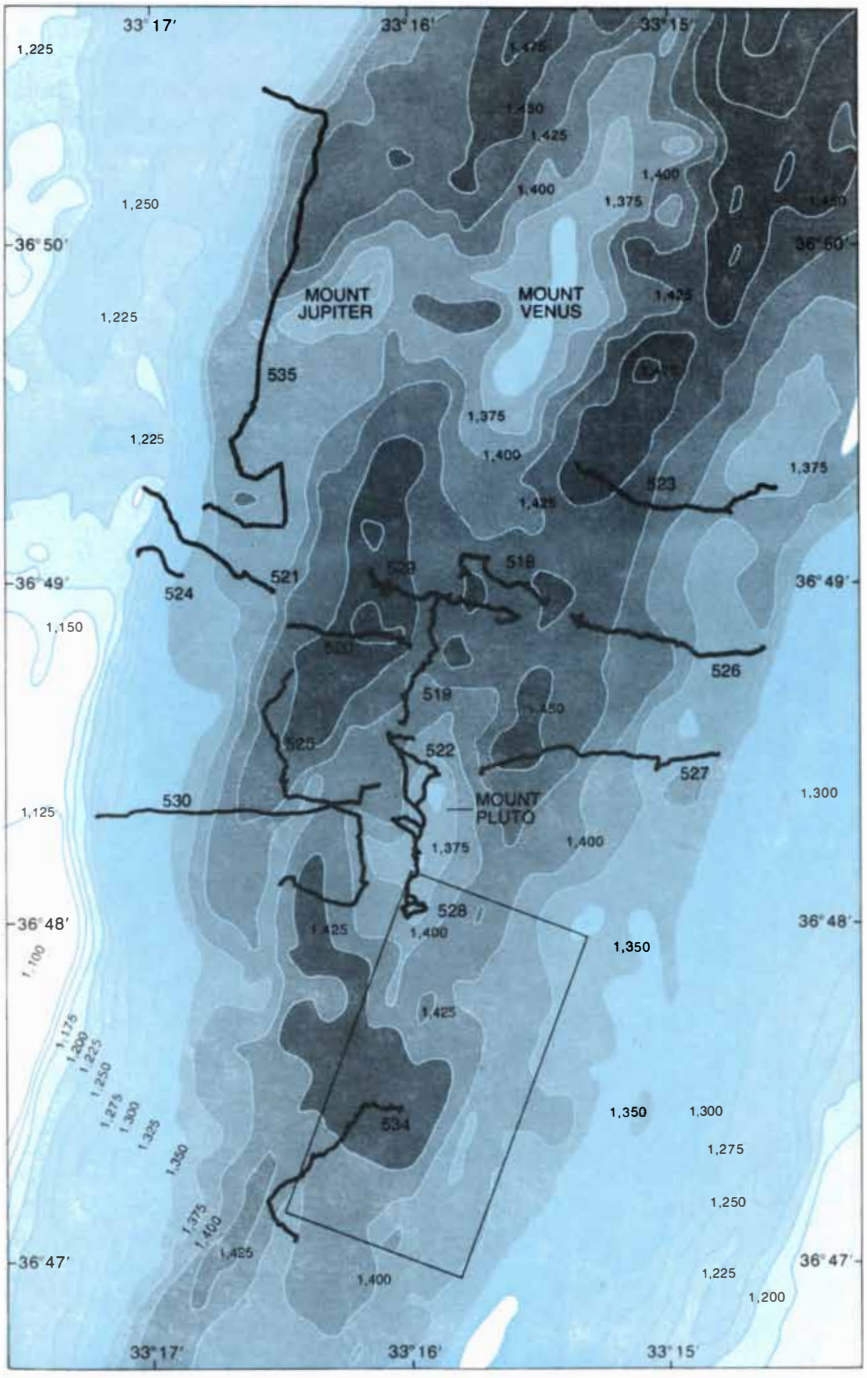
naturally present in the liquid volcanic rock become aligned with the earth's magnetic field as the molten rock freezes. Records of the fossilized magnetic polarity, obtained by magnetometers towed on the ocean surface, revealed symmetrical bands of periodic reversals of polarity on both sides of the mid-ocean ridges. Concurrent land-based studies of polarity reversals in precisely dated stratigraphic sections that included volcanic rocks provided a time scale for the global reversals in magnetic polarity [see "Sea-Floor Spreading," by J. R. Heirtzler; SCIENTIFIC AMERICAN, December, 1968]. The conclusion seemed incontestable: the volcanic sea-floor basement was spreading away from the mid-ocean-ridge systems, growing by constant additions of fresh volcanic material at the central rift, and in the process was recording the polarity of the earth's magnetic field as a function of time, rather like a giant tape recorder.

Beginning in 1968, deep-sea drilling by the *Glomar Challenger* confirmed the hypothesis of sea-floor spreading by showing that sediments overlying the volcanic basement became progressively older with distance east or west from the Mid-Atlantic Ridge. In mid-1970 the oldest volcanic basement rock yet recovered in the Atlantic was brought up in a core drilled by the *Glomar Challenger* at the base of the continental slope east of Cape Hatteras. At least 150 million years old, the sample exhibits all the mineralogical and chemical characteristics of present-day volcanic extrusions on the Mid-Atlantic Ridge. It appears that the rock was extruded at the Mid-Atlantic Ridge shortly after North America broke away from Africa at the beginning of the current episode of sea-floor spreading.



netized rock exhibits symmetrical bands of alternating polarity on both sides of the ridge. In the map at the right, which corresponds to the area within the rectangle in the map at the left, the numbers next to the colored lines that run parallel to the ridge axis represent the ages of the magnetized rocks in millions of years. (The contours are given in fathoms.) The region explored by Project FAMOUS lies within the rectangle in map at the right. A more detailed map of area explored by submersibles appears on next page.

Our primary purpose in Project FAMOUS would be to examine the details of the structure of the median valley in the Mid-Atlantic Ridge and to learn as much as possible about the extrusion and accretion of volcanic rock associated with the spreading process. We planned to begin with a series of broad surveys over the area of interest, gradually focusing more closely on the most promising area as both our knowledge and our technical capability improved. By the time the project was completed we had available for study the data collected by some 25 surface cruises, two aeromagnetic surveys and 47 coordinated submarine explorations of the rift valley.

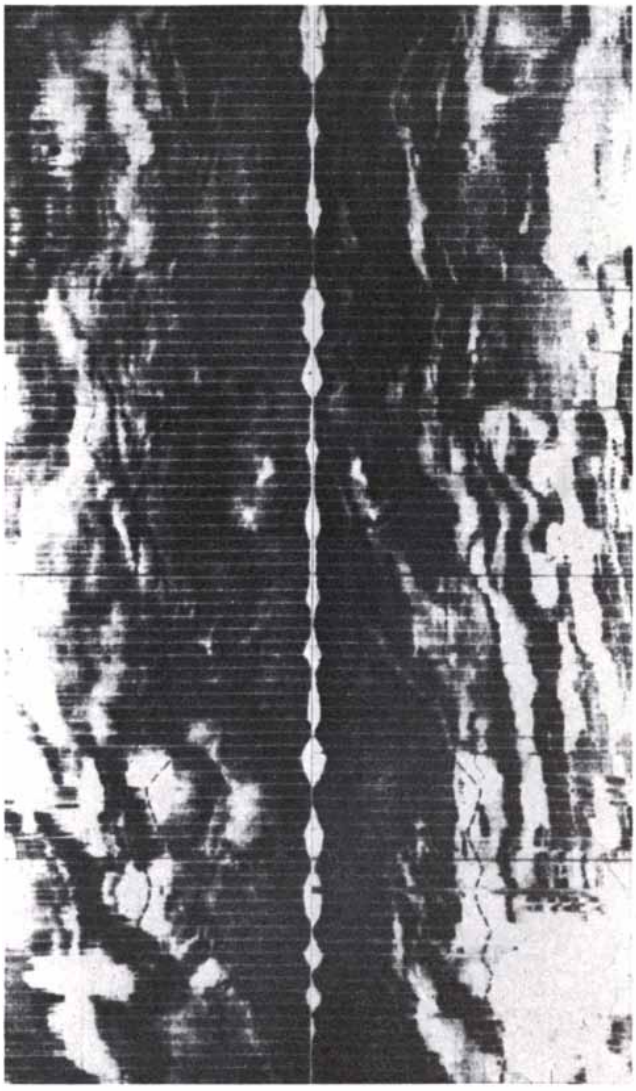


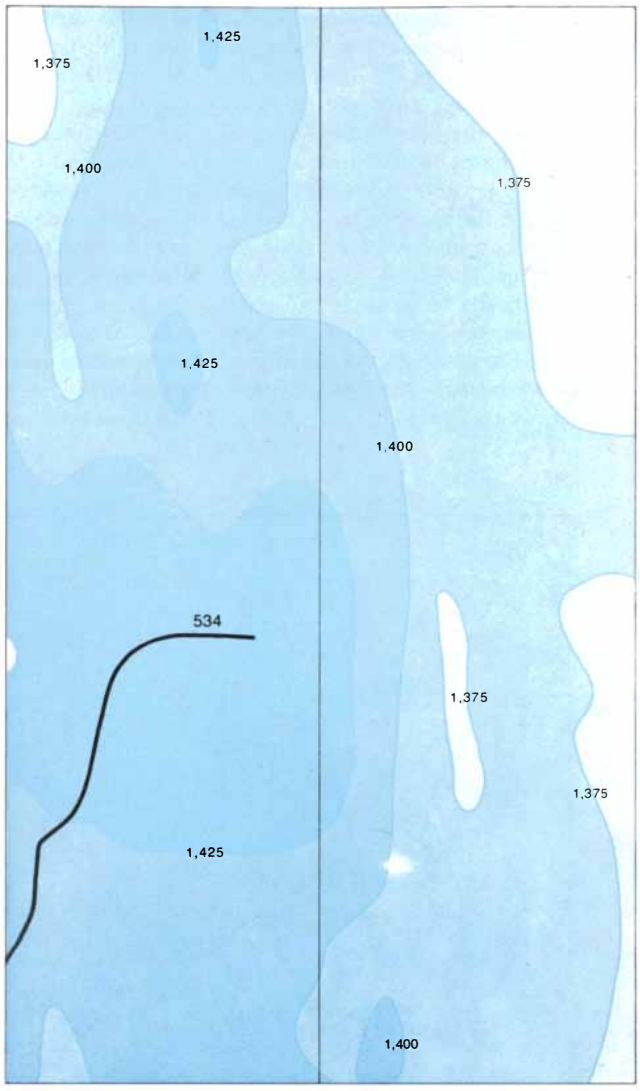
RIFT VALLEY explored in detail embraced an area approximately four kilometers by six kilometers. (Area inside rectangle appears in illustration on opposite page.) The rift valley is bounded on the west by a steep cliff, the West Wall, some 1,000 feet high, and on the east by a series of shallow terraces. The tracks of 15 exploratory dives made by the *Alvin* are depicted and numbered, beginning with dive No. 518. (The *Alvin* had previously made 517 dives elsewhere.) In addition the French submersibles *Cyana* and *Archimède* made dives in the rift-valley-and-fracture zone that lies to the north of Mount Venus in the Mid-Atlantic Rift. The *Cyana* made 12 dives in 1974; the *Archimède* made six in 1973 and 12 in 1974.

In the fall of 1971 a regional aeromagnetic survey established the existence of the typical pattern of magnetic anomalies on both sides of the axis of the mid-ocean ridge and indicated that the African and North American crustal plates were moving away from the axis at the rate of about 1.5 centimeters per year. Later in the year the first precisely navigated wide-beam acoustical survey revealed that the floor of the rift valley was fairly flat. Whereas the entire valley is about 30 kilometers wide, the inner floor is between one kilometer and two kilometers wide. This particular section of the Mid-Atlantic Ridge is offset by two fracture zones about 30 miles apart. The one to the north was designated Fracture Zone A, the one to the south Fracture Zone B [*see illustration on preceding two pages*].

Starting in the spring of 1972 the U.S. Naval Oceanographic Office began a narrow-beam echo-sounding survey of the area between the two fracture zones. The final product of the survey, completed early in 1973, was a set of bathymetric charts with a contour interval of five fathoms (about 10 meters). A different acoustical technique was simultaneously employed by British participants in the project. They used a side-scan instrument, embodied in a seven-ton submerged system, that directs its acoustic radiation at the sea bottom obliquely and thus can record echoes from many topographic features at different distances, all with one outgoing pulse. Since each pulse can irradiate large areas of the sea floor, it does not take long to assemble a regional echo map in which the major linear features of the area are clearly defined. The side-scan technique showed the steep west wall of the valley, the less steep east wall, the "corner hills" near Fracture Zone B and the hills down the center of the rift-valley floor, all in accurate relation to one another.

During the summer of 1972 the French research vessel *Charcot* carried out detailed bathymetry, a magnetic survey and a bottom-sampling program along the northern half of the rift valley south of Fracture Zone A. The survey showed the presence of a central hill on the valley floor (later to be known as Mount Venus) and confirmed that the magnetic-anomaly patterns were broader on the east than on the west, indicating that the spreading to the east is faster than that to the west. Rocks recovered





SIDE-SCAN ACOUSTICAL TECHNIQUE produced the images shown at the left. In this technique a "deep-towed fish," designed to operate 100 to 200 meters above the sea floor, directs its acoustic radiation obliquely at the bottom (*see illustration on next page*). Depending on the topography, each outgoing pulse can produce multiple echoes at differing distances from the axis along which the instrument is traveling. The two images that are reproduced here represent the left-hand and right-hand reflections from the sea floor depicted in the contour map of the same area at the right. Elevations in the floor produce reflections that appear bright in the acoustical record. It can be seen that in general the right (east) side of the record contains more linear features than the left (west) side.

from the rift valley popped and exploded on the deck of the ship, apparently as a result of the release of trapped gas. The rocks appeared to be very fresh, suggesting recent volcanic activity. A sled carrying cameras was towed along the bottom and obtained the first photographs of the bulbous and tubular lavas that were to become so familiar to the diving scientists.

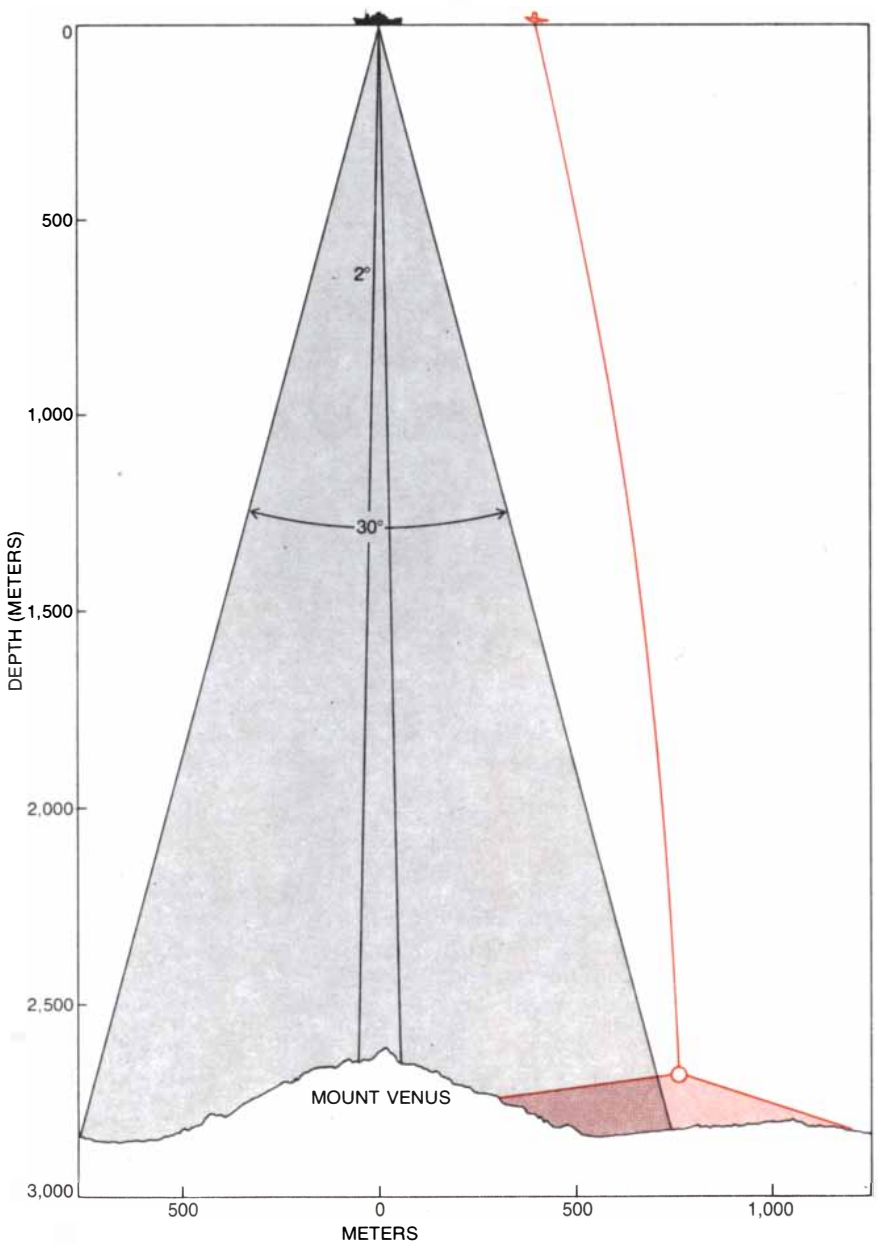
The French survey was supplemented and extended south to Fracture Zone B in late 1972 by the *Atlantis II* from Woods Hole. Three radar transponder beacons were anchored on mountain peaks in the rift to serve as navigation reference marks for a series of closely spaced bathymetric and magnetic survey lines. Our data confirmed the asymmetry in the median valley, with the western wall rising very steeply and the eastern wall rising more gradually in a series of steps.

In order to survey and sample this portion of the valley in detail we moored two acoustical transponder beacons on a terrace near the top of the west wall. Since the beacons were supported on floats 100 meters above the sea floor, they would not be affected by storms, unlike the radar buoys, which could swing in a large circle around their mooring. In response to a signal transmitted from the ship, the beacons transmit a characteristic signal, each identifiable by its own wavelength located between 10 and 14 kilohertz. The two-way travel time provides a precise measure of the slant range between the ship and the transponder beacon. The depth of each transponder and the distance between transponders is established by running a survey pattern over them; a shipboard computer is then used to generate a best estimate of the depths and base-line length. With these parameters fixed, the slant ranges can be converted to horizontal distances from points directly above the transponders, thereby uniquely fixing the ship's position in relation to

the base line. With a two-transponder base line it is necessary to know, of course, which side of the base line the ship is on, but this ambiguity is easily resolved from the known differences in bathymetry across the base line.

Our dredging and photographic surveys required the lowering of the dredge or the camera frame on a long cable behind the ship. The ship's position could now theoretically be plotted to an accuracy of a few tens of meters. (We now had to ask: For which part of the ship do we want a position?) The towed devices, however, would travel at some considerable but unknown depth and distance behind the ship. Thus a third transponder was attached to the cable close to the dredge or the camera. This transponder transmitted a signal that not only was received back at the ship but also triggered a second set of responses from the base-line transponders. The secondary base-line responses were received by the ship some seconds after the primary responses. Since the ship's position was continuously being fixed on the basis of the primary responses, the depth and position of the "relay" transponder on the cable could be computed. Any depth ambiguity in the solution could be resolved by entering into the calculation an estimated depth (based, for example, on the amount of cable paid out). Essentially the same system would be used later to navigate the manned submersibles, except that an automatic acoustic transmitter would be installed in the submarine to perform the same function as the towed relay transponder [*see top illustration on opposite page*].

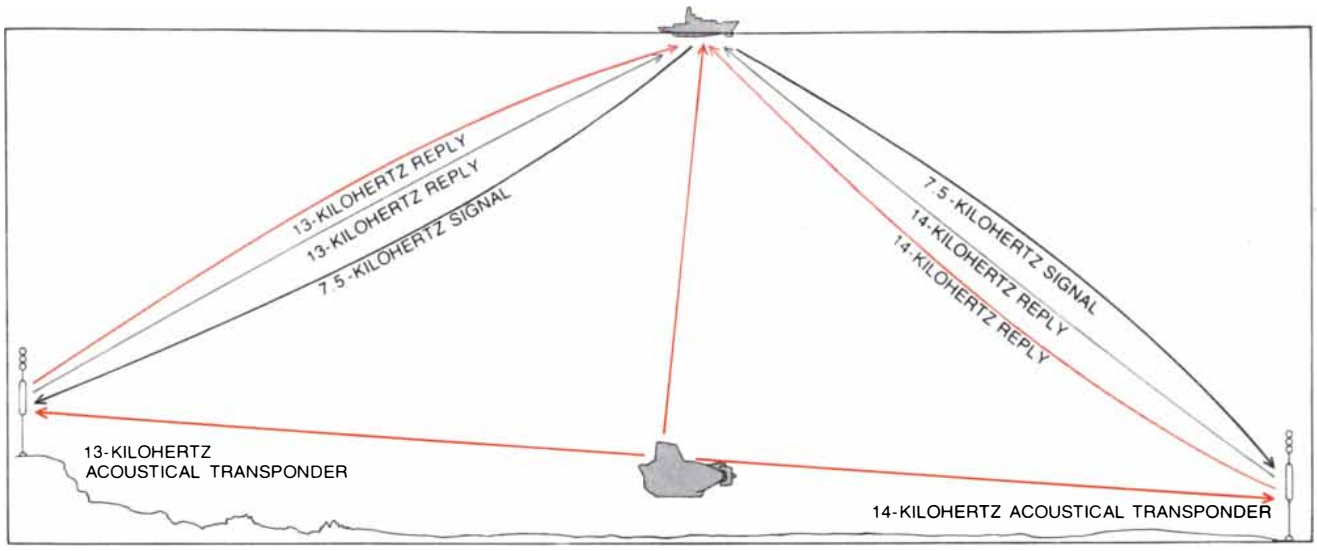
In practice all the raw data (acoustic travel times) were stored on magnetic tape for later replay at home, to allow the filtering of spurious positions introduced by acoustic noise or "bottom bounce." Our real-time output on shipboard was a tabletop *X-Y* plotter that drew a small dash to represent the ship's position and a small cross for the position of the towed instrument. On later cruises real-time tracking was provided on a fluorescent screen coupled to a thermal printer. Both systems produced inked or printed records of the intricate maneuvers executed during camera or dredge stations, with the positions of the ship and the instrument being recorded every 20 to 40 seconds. By laying plots of portions of our bathymetric maps on the printer, or on transparent overlays on the screen, we could maneuver the ship and instrument package to photograph or sample any given bathymetric feature. Since the time of each acoustic fix was known, and since the time was recorded on each photograph, the position of any given feature photographed on the sea floor could be established.



THREE ACOUSTICAL-MAPPING INSTRUMENTS supplement one another in recording the depth and the topography of the ocean floor. The conventional echo sounder, which has a beam angle of 30 degrees, provides quick, rough coverage of large areas. A newer surface instrument with a beam angle of only two degrees yields much finer detail. The deep-towed fish provides even finer detail of certain features, but its images require more interpretation.

We found that when the dredge touched bottom, vibration in the cable destroyed the acoustic signal; gaps in the track of the dredge indicated the positions of bottom contact and hence the location of any samples recovered. We also learned that we could reposition the dredge precisely in areas where rock had been revealed by photographs, and in these places we invariably recovered sizable quantities of rock fragments resembling those observed in the photographs. This procedure soon led to a system we called touchdown dredging; instead of dragging the dredge for half a mile or more across the bottom, as was the common traditional practice, we would maneuver it into position over a specific target, lower it to the bottom, hold it there for a few hundred yards of dragging and then bring it to the surface. In this way it was possible to relate recovered rock samples directly to specific bottom features.

Many of our photographs showed tubular or wormlike lava forms. Dredg-



POSITION OF THE ALVIN and the other submersibles during the Project FAMOUS dives was established with the aid of acoustical transponders anchored to the ocean floor, which provided an accurately known 4.5-kilometer base line for computing the travel times of acoustic signals. Three transponders laid out in a shallow triangle were actually used; only two of them are shown here. The surface vessel guiding the submersible established its own position by sending out a 7.5-kilohertz pulse and timing the replies from the anchored transponders at two different frequencies (13 and 14 kilohertz). The submersible also emitted pulses at 7.5 kilohertz that were received directly by the surface tender and that also triggered separate responses from the anchored transponders. The difference in arrival times between the direct signal from the submersible and transponded signals were used to compute submersible's position.

ing in those areas, we recovered samples with a circular or wedge-shaped cross section. Many of the samples could be reassembled on board ship, much as an archaeologist reassembles broken pottery. These reconstructions reproduced exactly the tubular forms observed in our photographs and further confirmed the fact that most of the material had indeed been recovered from a small area, in most cases representing only one or two distinct outcrops. Photographs made in Fracture Zone B showed angular, irregular or slabby rock fragments; again our dredging recovered similar material. Many of these samples had been crushed, sheared and recemented, as one might expect in an active geological fracture zone.

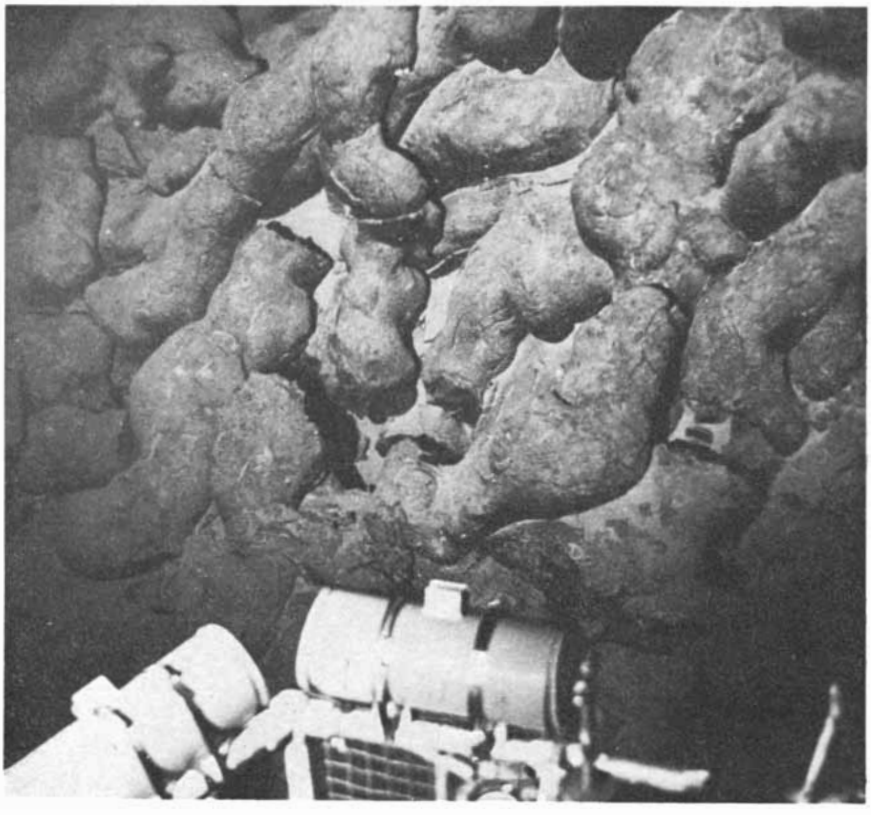
In the summer of 1973 the *Atlantis II* returned to complete the survey of Frac-

LAVA EXTRUSIONS in rift take exotic forms. The large toothpastelike extrusion in this photograph taken from the *Alvin* is typical. A remote manipulator is sampling an adjacent blisterlike extrusion. Photograph was made on dive No. 522 on Mount Pluto.

ture Zone B and of the next rift-valley segment to the south. Data from this survey would help us to make the final decision on the dive site and would extend our regional coverage to help put the detailed dive operations into proper perspective. The acoustical navigation system had been refined on the basis of the previous year's experience. We would also attempt to make detailed studies of heat flow from the bottom in and near the median valley, and we would attempt to monitor and compute the position of earthquake shocks in the median valley and the fracture zones. The navigation system now was programmed to make use of a network consisting of three transponders, which both extended the range of the system and removed the base-line ambiguity.

By working with this system to position the ship over known pockets of sediment in the median valley, we were able to drop sensitive thermocouple probes into the sediment. Then, measuring the temperature differential between two or more points on the probe in the sediment, and knowing the thermal conductivity of the sediment, we were able to compute the rate of heat flow. Such measurements were obtained within a few kilometers of the valley axis and around the active transform section of Fracture Zone B. The results showed low to normal values in the median valley and above-average values on the flanks of the valley and in the fracture zone. In the fracture zone values 10 to 12 times normal were observed. These results, along with veining and cementation in the fracture-zone rock samples, seemed to provide evidence of escaping hydrothermal solutions in or near the fracture zone.

The earthquake survey also made use of a version of the acoustical navigation tracking program, modified to plot the position of three sonobuoys simultaneously. Each sonobuoy is a small floating instrument carrying a radio transmitter and a suspended hydrophone. Three of these sonobuoys were set out in a triangular array that was constantly tracked until the sonobuoys drifted out of range. Earthquake shocks picked up by the hydrophones were transmitted to the ship by radio and were recorded on magnetic tape; the signals were also amplified and broadcast in real time over loudspeakers in the laboratory. Since the positions of the three sonobuoys were known, differences in the arrival times of the earthquake shocks at the three hydrophones could be used to triangulate the position and approximate depth of the shock, just as is done with an array of three or more land-based seismic stations.

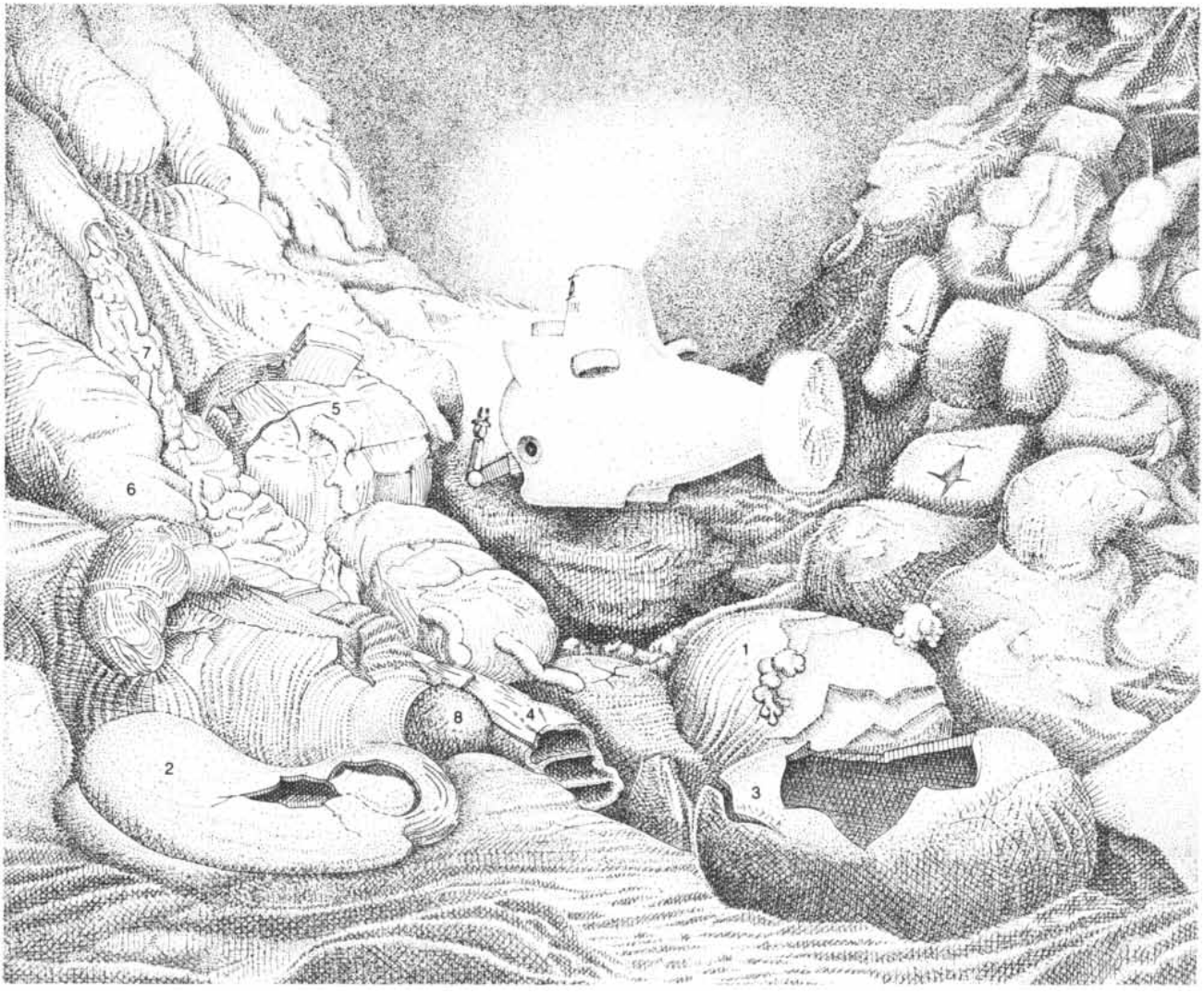


ELONGATED AND DRAPING LAVA PILLOWS, which are more characteristic of lava on land than they are of flows under water, appear in this photograph made on the *Alvin*'s dive No. 521 near base of West Wall of the rift. Canisters in foreground contain water samples.

What was to be one of our most exciting stations took place on a calm, starry night. With the sonobuoy array laid out, the *Atlantis II* moved slowly toward the steep west wall of the median valley as we lowered one of our new "Big Bertha" dredges. We planned to tow the dredge up the face of the wall. At this point, although we had begun to refer to the feature as a wall, we were aware of the way echo sounding tends to exaggerate such features, and we therefore assumed that the wall was actually no more than a steep slope, perhaps one of 30 to 40 degrees. As the dredge touched bottom at the base of the slope we began to receive ringing, banging sounds through the sonobuoy loudspeakers. When we slacked off on the cable, the noises stopped; as we took in the cable, they resumed. We watched the tension meter on the dredge cable expectantly as we approached the west wall. The echo sounder showed that the bottom was rising rapidly, when suddenly it disappeared entirely. We changed scales and tried our backup 12-kilohertz echo sounder. There was still no return. The bottom was so steep it could not reflect the signal!

Simultaneously the ringing of the sonobuoys was replaced by an ominous silence. The cable tension built up to 10,000, 12,000, then 15,000 pounds, more than twice its normal working range. Stopping the ship's engines, we drifted slowly backward, taking in cable carefully, trying to keep the tension below the 20,000 to 25,000 pounds that could break it. Finally we were directly back over the dredge, with bottom beginning to show again on the echo sounder. There was nothing more to do now but to take in the cable slowly, waiting for something to give. Suddenly there was a loud report from the loudspeakers, and the ship shuddered and reverberated as the tension meter swung wildly, then dropped back to the normal 5,000 to 6,000 pounds. We began reeling in cable cautiously, wondering when the frayed end would appear. A series of shocks and reverberations continued to roll in over the loudspeakers, although the dredge, if we still had it, should have been well clear of the bottom.

Finally the dredge appeared over the stern of the ship, with an immense freshly fractured hemisphere of pillow lava



BIZARRE SUBMARINE LANDSCAPE of lava extrusions was observed at the intersection of two lava-flow fronts in the Mid-Atlantic Rift. The drawing is based on a sketch of the scene made by one of the authors (Bryan), who served as one of the observers aboard the *Alvin* during its dives into the Mid-Atlantic Rift. The numbers identify some of the lava forms that were observed: (*1*) bulbous pillows with knobby budding, (*2*) a flattened pillow formed by the rapid drainage of lava while the skin was still plastic, (*3*) a hollow blister pillow formed by the drainage of lava after the skin had solidified, (*4*) a hollow layered lava tube formed by temporary halts in a falling lava level, (*5*) a bulbous pillow with a "trapdoor" and toothpaste budding, (*6*) an elongate pillow, typical of a lava extrusion on a steep slope, (*7*) a breccia cascade, formed on very steep slopes where the lower end of an elongate pillow has ruptured, releasing a cascade of fluid lava, and (*8*) an elongate pillow swelling into a bulbous form along a longitudinal spreading crack.

wedged in its jaws and several more large fragments caught in the chain bag. This was the largest single rock recovered during the project. It weighed about 400 pounds and had to be broken into three pieces for handling. The sonobuoys continued to record almost constant small shocks and vibrations for the rest of the night. We speculated that when the dredge broke loose, it triggered a series of rock slides on what must have been a very steep cliff. This speculation was given support the following year, when the submersible scientists observed piles of large talus blocks at the foot of the west wall. They also found many smaller piles of loose rock debris on terraces on the sides of the west wall, which turned out to be a spectacular cliff with a slope of nearly 80 degrees.

For a brief time the *Atlantis II* worked within sight of the French ships *Marcel le Bihan* and *Archimède,* as the first dives were being made near Mount Venus late in 1973. The initial dives showed that it was feasible to work with a submersible in this rugged terrain, to recover rock samples and to do geologic mapping. These preliminary submersible dives, before the major submersible effort planned for 1974, were of immense practical value in developing work and data-handling routines for the following year. For example, the *Archimède* detected bottom currents that were swifter than we had anticipated. A quickly instrumented program for metering ocean-bottom currents disclosed that the currents were nearly all tidal in nature and revealed no evidence of currents strong enough to be dangerous to submersibles.

A deep-towed instrument package, provided by the Scripps Institution of Oceanography, was brought into play at about the same time to study the microtopographic relief and the localized variations in the magnetic anomalies. This near-bottom survey confirmed the central magnetic and bathymetric asymmetry revealed by the surface-ship data,

although it suggested that an asymmetry had existed in the opposite direction a few million years ago.

Following closely on the *Atlantis II* cruise the U.S. Naval Research Laboratory introduced a major new bottom photographic technique in the same area. By suspending a strong stroboscopic-flash lamp high above the camera, they were able to illuminate and to photograph an area nearly 100 times as large as the area covered by the usual ocean-bottom camera. In fact, the area photographed with each flash of the system was approximately the same as the area covered by each ping of the narrow-beam echo sounder. Mosaics made from these photographs were later laid out to scale on a gymnasium floor to enable the diving scientists to preview the terrain in which they would work.

Throughout 1972 and 1973 the French submersible *Cyana* and the U.S. submersible *Alvin* were undergoing their refitting for operation at rift-valley depths, and the diving scientists and pilots were engaged in a training program so that they would be able to work effectively over the volcanic terrain of the valley floor. In contrast to the practice of the Apollo lunar program, all the men selected to observe the rift valley from submersibles were professional earth scientists, and several had had previous experience with the submarine and over volcanic terrain. Nevertheless, we would rely heavily on our pilots' understanding of the features we would be searching for. In Hawaii the scientists and pilots spent five very profitable days, accompanied by several French colleagues. There they observed a variety of flow forms and fractured lava flows both above and under water, watched an active volcanic spatter cone and visited an area where divers had recently sampled and photographed flowing lava under water. The French made other trips to Iceland and Africa to observe areas generally supposed to be dry-land extensions of oceanic-rift systems.

In the summer of 1974 the French submersibles *Cyana* and *Archimède,* respectively accompanied by their mother ships *Le Noroit* and *Marcel le Bihan,* and the U.S. submersible *Alvin,* accompanied by its tender *Lulu* and the research vessel *Knorr,* all engaged in a coordinated diving and exploration program in the rift valley and the adjacent fracture zones. During the summer the *Cyana* completed 14 dives, the *Archimède* 13 dives and the *Alvin* 17 dives. These dives, and the preliminary seven made by the *Archimède* the year before, were man's first direct observation of the rift-valley floor.

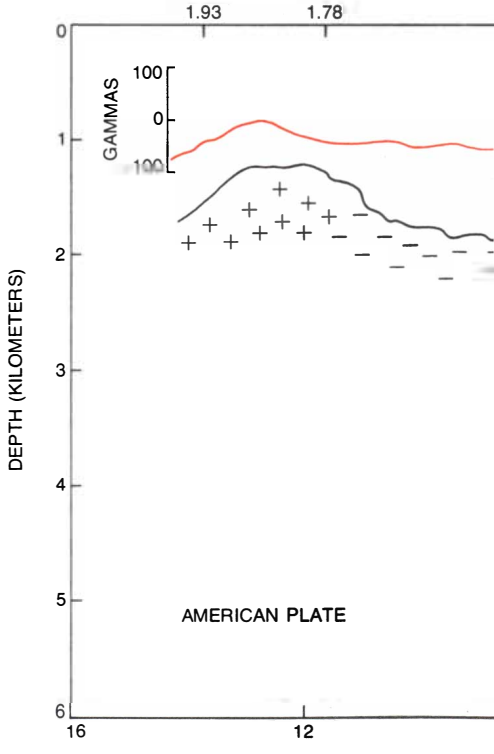
By good fortune it was possible to schedule the *Glomar Challenger* to drill in the area. The deep-sea drilling vessel was able to penetrate some 600 meters into basement rock 18 miles west of the dive area. The rocks collected by the submersibles and by the *Glomar Challenger,* and dredged by surface ships in the area in between, comprise one of the most remarkable sets of samples ever obtained from the sea floor. The shipboard and submersible samples together will make it possible to examine regional variations in composition over an area measuring some 30 by 60 miles. In addition the drill core provides an opportunity to look at vertical variations in the crust.

The *Archimède* began its diving program on and near Mount Venus, gradually working north toward Fracture Zone A, while the *Alvin* began diving near the next central hill to the south, which we called Mount Pluto. A secondary dive site for the *Alvin* was designated in Fracture Zone B. The *Cyana* was scheduled to begin dives in Fracture Zone A, but rough weather at the start of the program damaged the *Cyana* slightly and delayed the start of the *Alvin* operations for several days. The *Alvin*'s first few dives showed that visibility was excellent, that maneuvering was no problem and that sampling was as easy as we had hoped.

As the dives progressed the pilots found that it was possible to maintain almost continuous contact with the bottom. The scientists could recognize lava-flow fronts and conical structures named haystacks, which, because of their size and shape, were difficult to see in bottom photographs. The haystacks appeared to be small centers of lava extrusion. Major lava flows and vent structures were almost always confined to the central hills and were irregularly distributed along the central line of the valley, an observation strongly suggesting episodic volcanic activity. Minor lava flows and haystacks were observed on both sides of the central hills, particularly to the east. The submersibles collected precisely located and oriented rock samples, water samples and sediment cores. Almost continuous photographic coverage was obtained with semiautomatic cameras outside the hull of the vessels and both still and motion-picture cameras inside.

Small cracks in the sea floor near the axis of the mid-ocean ridge had been recognized first in photographs taken during the *Atlantis II* cruises. The extended-area bottom-camera system had clearly shown these features running for at least 100 meters and overlapping in an echelon fashion paralleling both the direction of the ridge and the major walls of the valley. The submersibles followed and crossed numerous fissures, measuring their heading, width, depth, longitudinal tilt, location and cross section. It was found that fissures exist everywhere from the valley's central line to its bounding walls on the east and the west, generally increasing in width with distance from the valley axis. The width varied from a few centimeters near the axis to tens of meters near the walls. Even the narrowest cracks were several meters deep. The wider ones were between 10 and 100 meters deep. In places there were small differences in elevation between one side of a fissure and the other. Across fissures up to a few meters in width it was possible to see matching halves of the same pillow lava on opposed walls.

Little sediment was found on the floor of the rift valley. Near the axis there was not even enough of it to half-bury pillow lavas a few tens of centimeters in diameter. Near the edges of the valley floor the sediment covered many such pillows



CROSS SECTION of rift is shown in relation to magnetic-field variations. Profile of the bottom is plotted from deep-towed

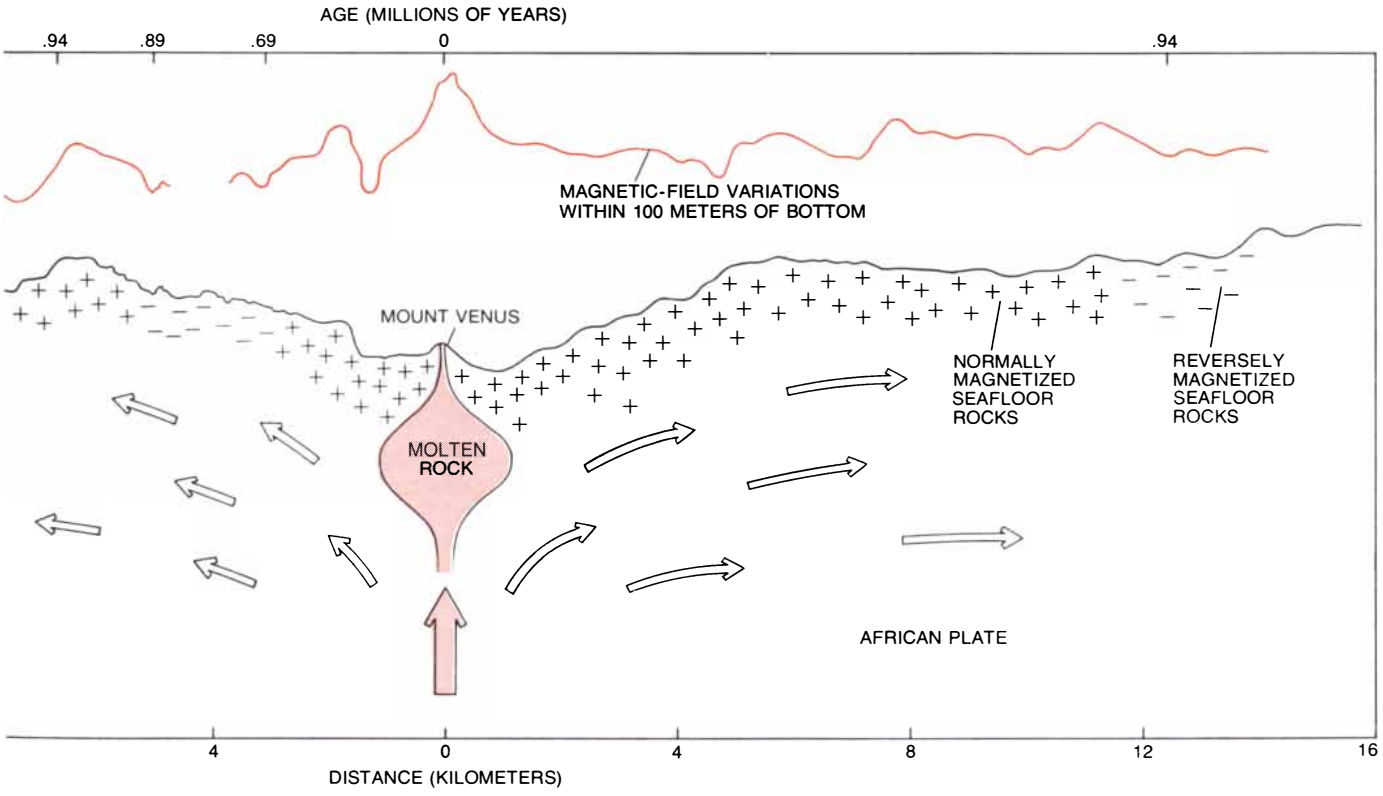
and possibly reached a depth of a meter or more. Dives in the fracture zones encountered a great deal of semiconsolidated sediment fractured by horizontal shearing. In those zones no fresh volcanic rock was found, even though a high heat flow had been recorded along one narrow fracture line. In Fracture Zone A scientists aboard the *Cyana* observed two manganese-covered areas that appeared to have been created by hot water flowing up through the sea floor. Although sensitive thermocouple probes were carried on the *Alvin* and also were towed near the bottom from surface ships, no significant thermal anomalies were found in the water.

Toward the end of 1974 we began the task of assembling the great mass of data accumulated during the cruises of the surface ships. Although much still remains to be done, some interesting patterns have begun to emerge. For example, by plotting the location and composition of rocks collected by acoustically positioned dredges in and near Fracture Zone B we have found that the sheared and altered rocks, which are typical of fracture zones, are correlated with a band of subdued magnetic anomalies over the fracture zone. The band follows the trend of the fracture zone and is nearly 10 miles wide. The earthquake surveys show that the narrow active transform fault is located near the north side of this band and is connected almost exactly with the limits of the active volcanic rifts in the center of the rift valleys north and south of the fracture zone. Apparently the active fault has shifted over the width of this zone; the volcanic rifts have probably moved north and south with the fault.

The bathymetric and magnetic asymmetry of the rift valley was confirmed in a striking way by the first analyses of the volcanic rocks. The analyses showed a regular variation in mineralogy and chemical composition with distance from the central volcanic hills toward the flanks of the valley. Moreover, the variations are more gradual to the east than they are to the west. In several dives the submersibles climbed the spectacular west wall, which rises some 300 meters in a series of closely spaced steps connected by nearly vertical fault scarps. No comparable feature was observed to the east; there, as the surface-ship data had suggested, the valley rises in a series of much wider steps, separated by short, steep slopes covered by broken fragments of lava.

Low areas flanking the central hills, which were originally thought to be grabens, or collapse depressions, were found to be simply low areas that had not quite been filled by lava flows converging on them from Mount Venus, Mount Pluto and the valley walls. True grabens were found at the base of the west wall and on the largest of the eastern fault scarps. These depressions suggest that the bases of the walls have pulled away from the valley floor, which has collapsed downward and inward toward the center of the valley. Indeed, one general impression is that the valley is near the end of a major period of structural extension and collapse and is just beginning to be inundated by new volcanic extrusions, represented by Mount Venus and Mount Pluto. Such episodes of alternating volcanic activity and quiescence may prove to be typical of slow centers of sea-floor spreading such as the Mid-Atlantic Ridge, where the rate of spreading may not be high enough to keep a rift constantly open to the underlying magma chamber that is the source of the lava flows.

No lava flows in actual progress were



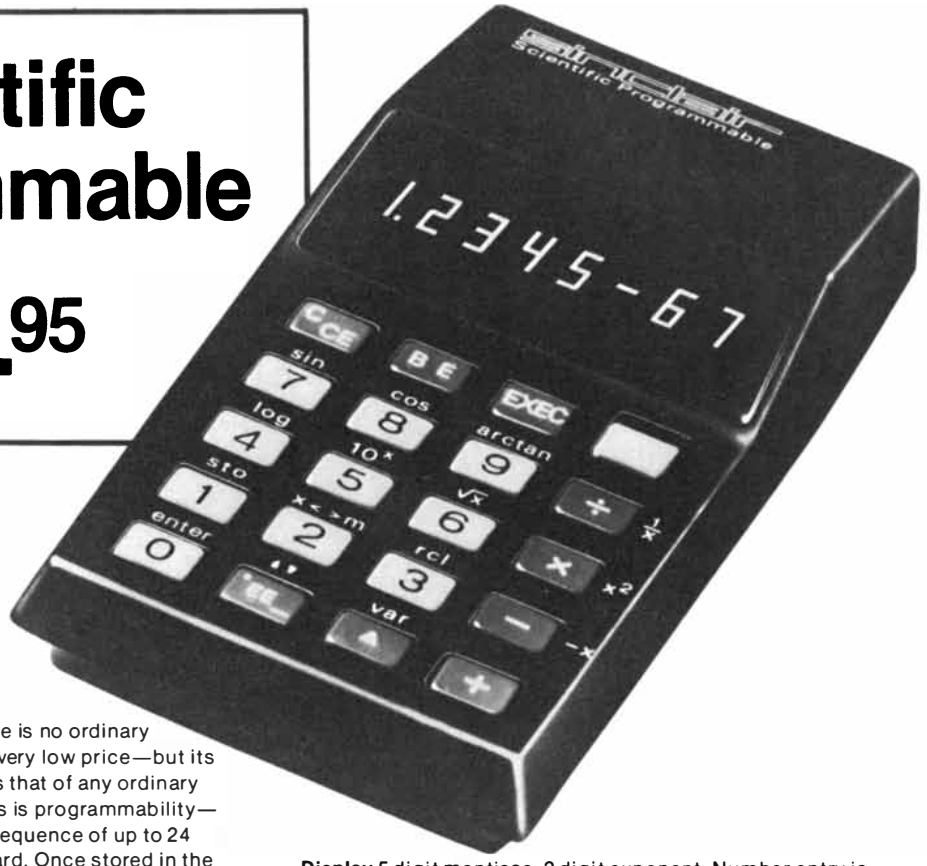
acoustical measurements. Rock of the floor is not symmetrically magnetized as measured from the center line, nor is it magnetized everywhere with equal strength. Ages of rock as identified from magnetic profile indicate that over the past 940,000 years average rate of spreading to the east has been about twice as fast as rate to the west: 1.3 centimeters per year compared with .74 centimeter.

observed by either the American or the French diving groups, although we suspect that the central volcanic hills may be very young; it is perhaps only a few hundred years since the lava last flowed out of them. Preliminary dating suggests that some of the lavas on the flanks of the valley may be much younger than their distance from the valley center would suggest, indicating that they must have emerged on the sides of the valley rather than in its center. The presence of these young flank lavas, together with extensive faulting and mechanical rotation of at least the upper 10 meters of the lava flows on the valley floor, presents difficult problems for the interpretation of the magnetic-anomaly pattern, which seems to persist in spite of such complications. It was formerly supposed that the magnetic patterns had to do with only the uppermost few tens of meters in the lava. That view may now have to be revised. The problem is further complicated by analyses of rock samples in cores taken by the *Glomar Challenger* at its drill site adjacent to the rift valley. The magnetic polarity measured in the rock samples does not seem to agree with the integrated polarity measured by surface-towed instruments. Preliminary results of magnetic-intensity measurements on samples collected by the *Alvin* also do not exhibit any simple relation to the magnetic intensity measured by deep-towed or surface-towed instruments. It now seems possible that the magnetic-anomaly patterns are related to intrusions more deep-seated than the lava flows. These deeper intrusions must be added to the diverging plates in a more regular and consistent way than the surface flows.

Many of the interpretations summarized here will necessarily be modified as the data are further refined and new facts come to light. Moreover, it may be that not all mid-ocean ridges, and probably not all parts of the Mid-Atlantic Ridge, have the same features as those of the area we have studied. Indeed, we expect that fast-spreading ridges, such as the ridge in the eastern Pacific, will turn out to be quite different. Perhaps most important, we have been able to demonstrate that the kind of fine-scale features that eluded detection in the unprecedented concentration of surface-ship studies pursued in the Project FAMOUS area can be examined routinely by manned submersibles. Further investigations of this kind are certain to add new dimensions to our understanding of processes operating on and below the deep-sea floor.

Scientific Programmable

$79.95



Sinclair's new Scientific Programmable is no ordinary calculator. It has only 19 keys—and a very low price—but its problem-solving capability far exceeds that of any ordinary scientific calculator. Sinclair's big plus is programmability—the ability to remember a calculation sequence of up to 24 steps entered directly from the keyboard. Once stored in the program memory, a calculation sequence can be recalled at the touch of a single key, and applied to new numbers to produce new results. So if you carry out repetitive calculations, the Scientific Programmable will save you key strokes and lots of time. It will also save you from errors, since the calculation sequence will be recalled exactly as you entered it, over and over again.

Programming Entering a calculation sequence into the program memory couldn't be easier. Press the BE key to tell the calculator to remember a sequence. Then key in the calculation almost exactly as you normally would. Use VAR at the points where you'll want the program to stop, so that you can enter new numbers or display partial results. During entry, the number of steps is displayed, so you won't exceed the program memory's capacity. When you've finished, press BE again to tell the calculator that the sequence is complete. The program is now available until you overwrite it with another program or until you switch the calculator off. Just enter new numbers, press EXEC, and get new results. Of course, the Scientific Programmable can be used as an ordinary scientific calculator. Even half-way through the execution of a program, you can stop, carry out a calculation from the keyboard, then press EXEC to continue execution of the program when you're ready.

Sinclair Program Library You don't have to be a programmer to use the Scientific Programmable's full problem-solving capability. Each calculator comes with a fully documented library of hundreds of programs to solve standard problems, and complete instructions on how to use them.

Applications The Scientific Programmable has many applications in areas of computation other than repetitive calculations. It can be used for the analysis of experimental data, the evaluation of integer functions, and with a variety of methods for the iterative solution of equations. They are dealt with in detail in the instruction book and Program Library. And, although the Scientific Programmable is no toy, it does play games—and wins!

Display 5 digit mantissa, 2 digit exponent. Number entry is floating decimal point and/or scientific notation; results in scientific notation.

Logic Reverse Polish Notation.

Functions Sine, cosine, arctangent (radians), log, alog (base 10), square root, reciprocal, change sign, clear/clear entry.

Memory Store, recall, exchange.

The Scientific Programmable is small enough to hold in your hand and big enough to use on your desk. It measures 6″ x 2-7/8″ x 1-1/4″, has non-slip rubber feet and a big green display. It uses a small, inexpensive 9v battery or the Sinclair AC adapter.

$79.95 includes battery, AC adapter, carrying case, instruction book and the Sinclair Program Library.

10 Day Trial Offer If the Scientific Programmable doesn't meet your requirements return it to us within 10 days for a prompt refund.

Warranty The Scientific Programmable is backed by Sinclair's full one-year parts and labor warranty.

sinclair™ **leaders in innovation.**

TO: **Sinclair Radionics Inc.**
375 Park Avenue, New York, NY 10022

Please send me ____ Scientific Programmable calculators.
I enclose a check or money order for _____.

Name _____

Address _____

_____

N.Y. residents please add sales tax.

AREA OF CENTRAL PLAZA at Cahokia is seen in this aerial photograph. The view is to the north; the shadows cast by the late afternoon sun accentuate the relief. The conical mound at left is No. 57; the platform mound at right is No. 60. Both were once enclosed by the timber wall that surrounded the plaza and Monks Mound, the 100-foot-high earthwork at upper right that dominated Cahokia.

A Pre-Columbian Urban Center on the Mississippi

About A.D. 1000 there arose in the area north and south of what is now St. Louis the most populous Indian settlements north of Mexico. Foremost among them was Cahokia, which included some 120 mounds

by Melvin L. Fowler

One of the largest earthworks built by ancient man anywhere in the world rises in the U.S. Middle West not far from where the Illinois and Missouri rivers join the Mississippi. Relatively few people other than prehistorians are aware of this colossal monument, and even prehistorians have only recently learned that it marks the center of a 125-square-mile area that contained the most populous pre-Columbian settlements in the New World north of Mexico. Today the area, which includes a floodplain, alluvial terraces and low bluffs along the east bank of the Mississippi north and south of St. Louis, is called American Bottoms. The huge earthwork at its center, a few miles east-northeast of East St. Louis, Ill., is known as Monks Mound. The aggregation of some 120 lesser earthworks that surround it is called Cahokia: the name of an Indian group living in the area at the time of the French colonization early in the 18th century.

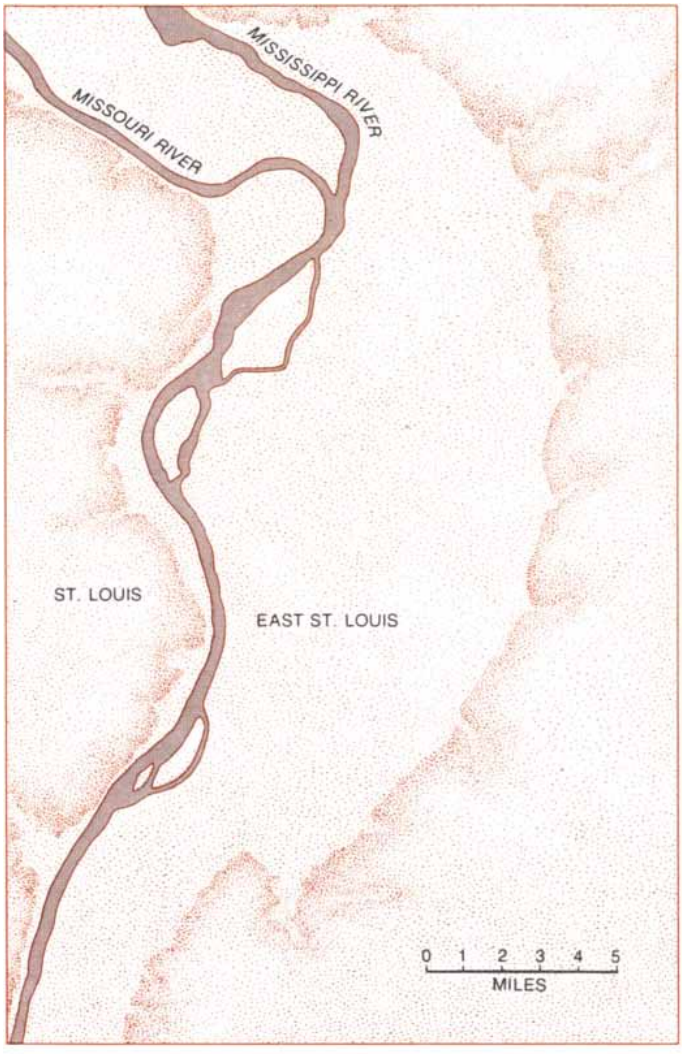
Monks Mound, which got its name from a short-lived Trappist settlement, is still an impressive affair [*see illustration on opposite page*]. Its base, which rises from a plain lying 417 feet above sea level, measures 1,000 feet from north to south and more than 700 feet from east to west, covering an area of about 15 acres. Its volume is estimated to be 22 million cubic feet; in North America only the Pyramid of the Sun at Teotihuacán and the great pyramid at Cholula are larger. It rises in four steps to a maximum height of 100 feet above the plain. The first terrace, occupying about a fourth of the surface area at that level, is 40 feet high. Three lobelike protrusions in the northwest quadrant of the mound collectively form the second terrace, which is some 62 feet high. The third and fourth terraces occupy the northeast quadrant; the fourth terrace, at an elevation of 100 feet, is three feet higher than the third.

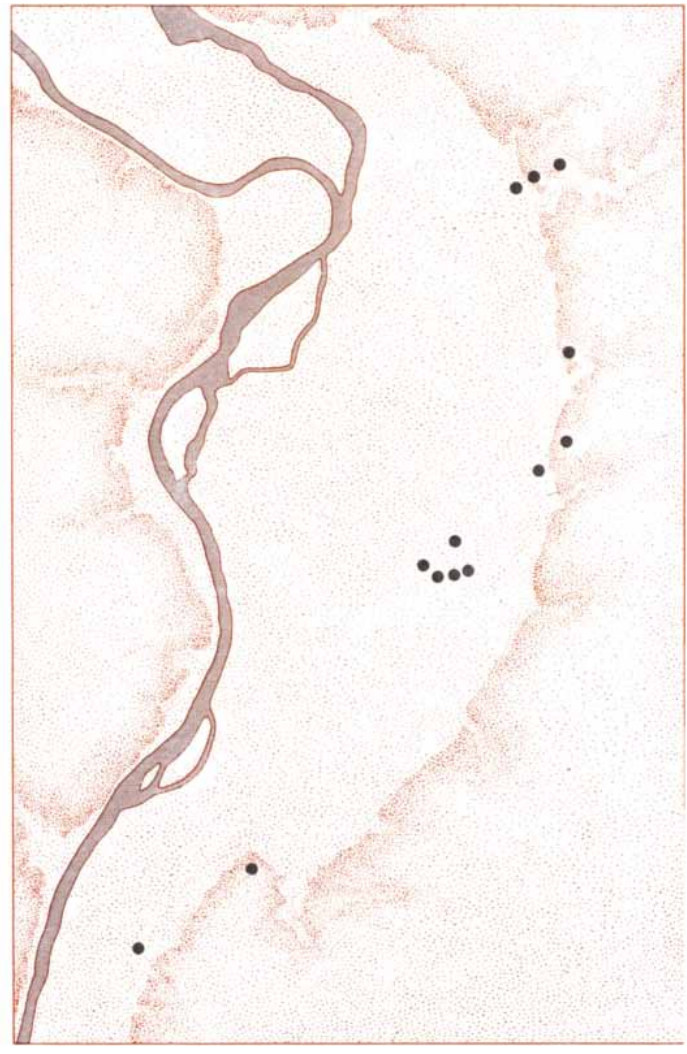
For all its impressiveness, Monks Mound is only a part of the even more impressive Cahokia group, and Cahokia in turn is only one, albeit the largest, of 10 large and small population centers and 50-odd farming villages that flourished in American Bottoms about the start of the second millennium (A.D. 1000).

How did these North American population centers arise? The best answer at present is that a complex feedback interaction involving population growth and an advance in agricultural productivity was responsible. Sometime late in the eighth century, it appears, the hoe replaced the digging stick in maize agriculture and a variety of maize became available that was better suited to the climate of the Middle West than the southern varieties that had been grown there up to that time.

Archaeologists, both amateur and professional, have worked in American Bottoms for a century, and in recent decades some of the sites that lacked official protection have been partially excavated before local construction destroyed them. Since 1950 at Cahokia alone individual or joint efforts by five universities and at least two museums have resulted in soundings and excavations at a score of localities. Yet only two or three of the 120 Cahokia mounds have been adequately excavated. Meanwhile the mounds that marked prehistoric popula-

CLOSE-UP OF MONKS MOUND, looking to the northwest, gives a sense of its dimensions: 1,000 feet at the base from north to south, more than 700 feet from east to west and 100 feet above the plain at the highest point. The mound is the largest pre-Columbian structure in the New World north of Mexico; it was built in successive stages between A.D. 900 and 1250.





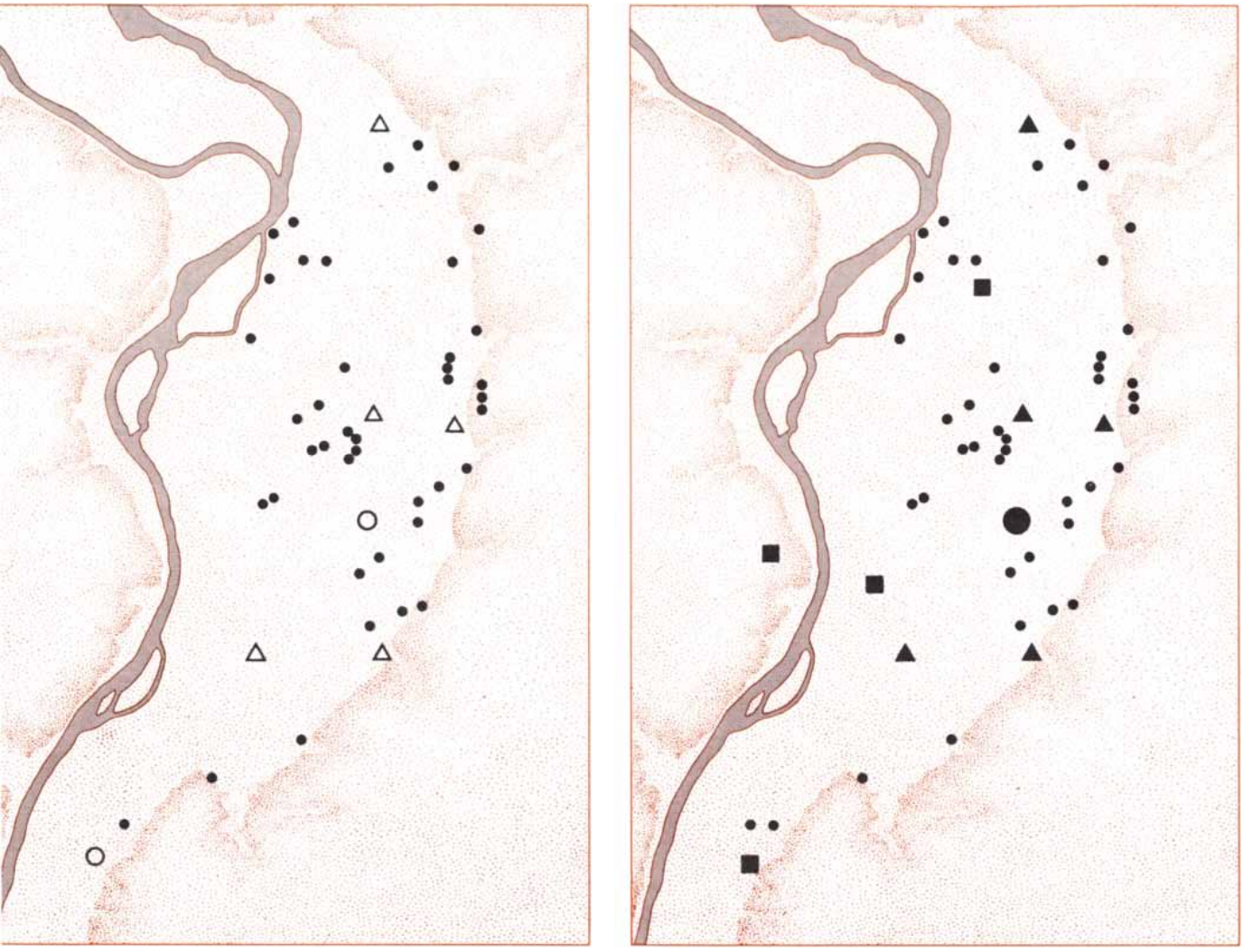
LINE OF BLUFFS (*left*) on both sides of a low floodplain outlines the American Bottoms region at and below the confluence of the Mississippi and Missouri rivers. The river channels shown are the present ones; the numerous lakes, creeks and sloughs occupying much of the bottomland have been omitted. About A.D. 800 (*right*) there were at least 13 Late Woodland habitation sites (*black dots*) in the region. Eight have been recognized on or near the upland bluffs; five, including one where Monks Mound would later rise, have so far been located in the richer bottomland. Farm productivity increased in American Bottoms soon thereafter.

tion centers of equal interest in St. Louis and East St. Louis have been obliterated by urban expansion. Nevertheless, it is possible, thanks to the work of recent years, to trace the evolution of the pre-Columbian settlements in American Bottoms with some degree of confidence.

The sophisticated kind of social and economic organization that is reflected by the construction of earthworks in American Bottoms was not the first to arise in the region. For several hundred years before that people belonging to a widespread culture, known as the Hopewell culture after a site in Ohio, had undertaken the construction of scores of earth enclosures, effigy mounds and other earthworks all through the Middle West. Another feature of the Hopewell culture was a widespread trade in exotic materials from as far away as the Rocky Mountains. Hopewell was a flowering of what is called the Middle Woodland period in this part of America. After A.D. 500, however, this cultural integration of a large area came to an end. The farming hamlets that thrived in the vicinity of American Bottoms about A.D. 800 were representative of the subsequent Late Woodland period.

The initial phase of the settlement at American Bottoms that led to the rise of Cahokia extended from A.D. 600 to 800. It is called the Patrick phase after A. J. R. Patrick, a physician of Belleville, Ill., who was a pioneer investigator of Cahokia in the latter half of the 19th century. Archaeologists from the University of Illinois at Urbana have excavated Patrick-phase pit deposits at a site at the western extremity of the Cahokia settlement, and our group from the University of Wisconsin at Milwaukee found Patrick-phase materials in the bottom occupation levels we uncovered below the eastern margin of Monks Mound. Other Patrick-phase remains have been found outside the Cahokia area.

The people of the Patrick phase appear to have been newcomers to American Bottoms, perhaps attracted there from the more typical river-bluff and upland farms by the fertile sandy loams that form the natural terraces and levees of the river valley. The information now available about the Patrick phase and an unnamed century-long phase that followed it is largely obtained from the analysis of pottery fragments gathered by surface collecting and the excavation of refuse heaps. Patrick-phase house sites indicate that residences at this time were rectangular and that the house posts were set in pits. There is little evidence of mound building between 600



FAIRMOUNT PHASE at Cahokia (*left*), which extended from A.D. 900 to 1050, brought many changes to American Bottoms. Two sites, Cahokia and Lunsford-Pulcher (*open circles*), came to include a number of earth mounds. Farm villages (*dots*) increased in number from 13 to 42. In addition five sites larger than villages arose, four of them near Cahokia (*triangles*); each included at least one platform mound. In the following 200 years (*right*), embracing the Stirling and Moorehead phases at Cahokia, other sites with plazas and platform mounds arose (*squares*). During this time Cahokia (*solid circle*) became the largest community in American Bottoms.

and 900, nor is it clear just how, sometime about 900, people of the Late Woodland period here adopted what American archaeologists call Mississippian culture.

The occupation phase that followed, called Fairmount, continued for 150 years. It was typically Mississippian in culture and includes ample evidence, in the form of mound construction and elaborate burials, that sharp social stratification and a centralized control of resources had arisen among the inhabitants of American Bottoms. Soil cores taken by investigators from Washington University indicate that the first work at Monks Mound probably began at this time. My group's excavation of a small "ridgetop" earthwork, Mound 72 near the middle of the Cahokia group, produced indications that the builders of other mounds in the vicinity followed a plan calling for the overall orientation of the settlement along a north-south axis.

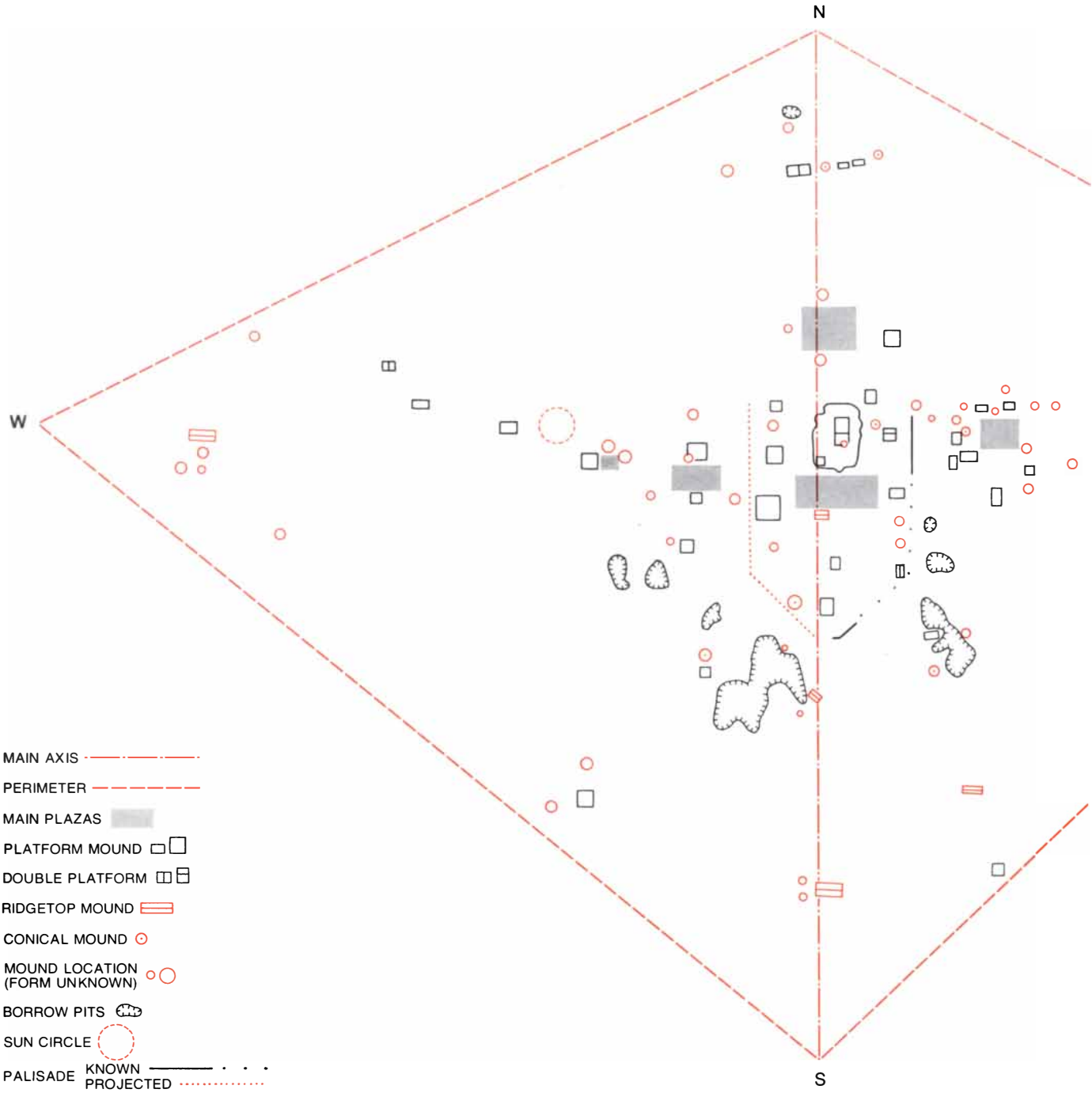
In view of the enormous quantities of fill required for the construction of the Cahokia earthworks, one immediate question is where did the soil come from? The answer is that the builders followed a procedure still used today: they dug "borrow pits." So far nine borrow pits have been located at Cahokia. The largest, about 800 yards southwest of Monks Mound, covered a 17-acre area and was about six feet deep. The second-largest pit covered nearly eight acres and the others ranged from about two acres to less than one acre. A group from the University of Illinois at Urbana has investigated one of the borrow pits that probably supplied the earth fill for an early phase in the construction of Monks Mound. The pit had later served as a trash dump; when it was full, the area was leveled and a large platform mound was built on top of the former excavation.

The builders did some of their digging with tools made from the hard, fine-grained stone known as chert. Quite probably they also worked with wood tools that have long since disintegrated, and they carried the earth fill from the pits to the construction sites in baskets.

Among the earthworks at Cahokia, Monks Mound is unique in shape. Mound 72 is one of six known ridgetop mounds in the group. By far the most common mound shape at Cahokia, however, is the platform mound; 28 square, oblong or oval single-platform mounds and four stepped, or double-platform, mounds have been identified [*see illustration on next two pages*]. In some instances excavation has shown that wood

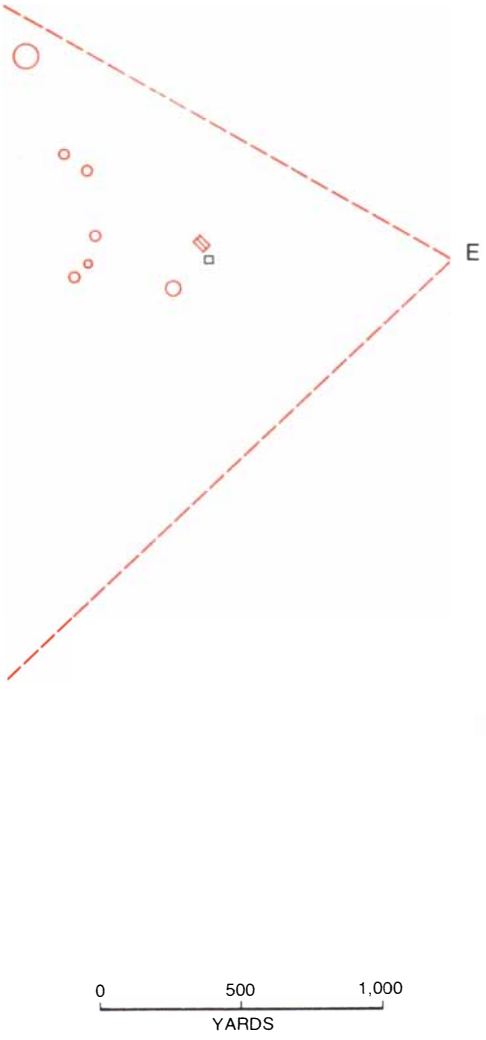
structures were built on the tops of the platform mounds, and so it is generally assumed that all the platform mounds served as building sites. The double-platform mounds presumably were used for structures more important than those on the mounds with only one platform. Just which platforms supported ceremonial buildings and which were residential sites occupied by the Cahokia elite is a question that only additional excavation can answer. A fourth kind of mound was also built; it is conical. There were seven of these mounds at Cahokia, and their shape makes it improbable that they were used as building foundations. They may have been used as burial mounds.

Other mound-building communities arose in American Bottoms during the Fairmount phase. One of them, the Lunsford-Pulcher site some 10 miles south of Cahokia, may even have approached Cahokia in size and importance at the time. Five other sites, four



CAHOKIA MOUNDS once extended three miles from the East Group (*right*) to the Powell Group (*left*), both now destroyed, and 2.25 miles from the Kunneman Group (*top*) to the Rattlesnake Group (*bottom*). When the four cardinal points are bounded (*broken colored line*), the area enclosed is some five square miles. Many of the 120 or so mounds at Cahokia have been obliterated by plowing or construction. Only 92 appear on this plan, and the original shape of 47 of them can no longer be determined. The remaining 45 fall into four classes. There are 28 single platforms and four double platforms; all 32 probably had buildings on them. Seven mounds are conical and six are classified as "ridgetop." Four of the ridgetop mounds may have been intended to mark the two axes of the site. Open colored circles locate the 47 mounds of unknown shape; irregular black areas mark eight of the nine known "bor-

of them quite near Cahokia, are characterized by the construction of at least one platform mound. Elsewhere in the bottomland the number of farming hamlets and villages, all of them without mounds, had increased from a known total of 13 during the Patrick phase to more than three times that number [*see illustration on page 95*]. Two carbon-14 dates, both based on wood samples excavated at Cahokia, have helped to define the duration of the Fairmount phase. A house site excavated by workers from the University of Illinois at Urbana produced a carbon-14 age reading in the range between A.D. 685 and 985, and wood recovered from a ceremonial posthole underlying Mound 72 produced a reading in the range between A.D. 925 and 1035.



row pits" that furnished earth to the builders. A broken colored circle locates the Cahokia woodhenge. Shaded areas outline five possible main plazas. The perimeter palisade surrounding the center of the site has been traced only in part; a dotted line (*color*) suggests its possible further extension.

The most intriguing information unearthed so far concerning the intellectual and societal complexity of Mississippian culture about the end of the first millennium is associated with the Fairmount phase. In 1961 archaeologists from the Illinois State Museum undertook salvage archaeology in advance of Federal highway construction at Cahokia. They came on a series of soil stains some 900 yards west of Monks Mound. Tracing the stains, they found that many large upright timbers had once been arrayed in a circle. Within the circular enclosure the inhabitants of Cahokia were probably able, by sighting along certain marker posts, to observe the annual sequence of the solstices and the equinoxes. Solar observations of this kind, of course, can be the basis of a useful agricultural calendar. Several of these woodhenges were built in the same part of Cahokia over the centuries, but the first was built during the Fairmount phase.

Our group from Milwaukee had been struck by the fact that three of the six ridgetop mounds at Cahokia were located respectively at the eastern, southern and western extremities of the settlement. A fourth ridgetop mound, Mound 72, was located at the edge of the largest borrow pit, 800 yards south of Monks Mound. We found the position of Mound 72 suggestive. If one draws a line from Mound 72 to a ridgetop mound in the Rattlesnake group, some 600 yards farther south, a northward extension of the line crosses the southwest corner of the first terrace of Monks Mound. It seemed plausible to us that Mound 72 had served to mark a carefully calculated north-south center line at Cahokia.

To test our hypothesis we excavated a trench at the point in Mound 72 where, according to our prediction, the north-south center line would have crossed the structure. We uncovered a pit that extended well below ground level. At the bottom of the pit we found the impression of the butt of a very large upright timber. The timber itself was gone; perhaps the Fairmount-phase inhabitants had later removed it for use elsewhere. Still present, however, were the remains of logs that had been used to hold the large timber in position. A sample of this wood provided the carbon-14 date mentioned above.

As we continued to excavate Mound 72 it became clear that at the time the marker post was first erected, and continuously thereafter, the locale had served as a burial ground. It was no ordinary cemetery but one evidently reserved for the burial of the elite, perhaps even several generations of the same elite family. It was many years before the mound grew to its final ridgetop shape. We could recognize a succession of building phases; they included no fewer than six separate episodes of burial involving a total of at least 200 individuals.

The first episode, which was probably contemporaneous with the raising of the timber upright, began with the construction of a timber building on level ground. The lack of refuse in association with the building suggests that it was used as a mortuary, or charnel house. Early accounts of European contacts with some Indian groups in the Southeast include descriptions of the storage of the dead in such buildings and of their final burial only at a time that was ceremonially determined. In any event the building was eventually dismantled, and a mound was raised over a group of burials that included the body of one individual who had just died and the bundled bones of several others who had been dead for a long time. This kind of interment is suggestive of a ritual that postpones the burial of lesser kin until a kinsman of high status dies, whereupon all are buried as a group.

The earthwork that was built over this first mass burial was a small platform mound. Nothing was added to the mound for some time, but one pit was dug on the east side and another on the south to accommodate group burials. These pit burials comprised the second episode of burial at the site.

The third episode involved two further excavations and additional construction. One of the excavations was a pit that intruded into the fill covering the pit dug earlier on the south side of the mound. No bodies were buried in the new hole; instead the diggers placed pottery, shell beads and projectile points in the pit and then refilled it. At about the same time a rectangular pit was dug at the southeast corner of the mound and 24 individuals were buried there. Work did not stop with the refilling of the pit but was continued until a mound was raised above it that extended the initial platform mound to the southeast.

Just as the erection of a north-south

marker post suggests sophisticated planning, so the fourth episode of burial at Mound 72 indicates that the social system of Fairmount-phase Cahokia was a distinctly stratified one. First a large pit was excavated about 10 yards southeast of the extension that had been added to the platform mound during the previous episode. Between the pit and the mound extension a small earth mound was raised, and the bodies of four men, with head and hands missing, were placed on top of it. In the pit the bodies of more than 50 young women, all between the ages of 18 and 23, were placed side by side. Finally earth was heaped over both group burials so that the platform mound was extended still farther to the southeast.

Although there is no physical evidence that the young women met a violent death, their closeness in age argues strongly against their having died from disease or from some common disaster. It is difficult to avoid the conclusion that, like the four mutilated men, the women were sacrificed, probably as a part of some kind of funeral ritual.

On whose behalf was the sacrifice made? Not far from the pit, at the place where the timber upright had once stood, we uncovered another group burial. It contained the remains of an individual of obvious importance who had been buried soon after his death. His body had been placed on a platform made up of thousands of shell beads; nearby were a number of bone bundles and the partly disarticulated remains of other individuals.

TAPERED BEAKER from Cahokia is some nine inches high. Its spiral scroll design is more sophisticated than most of the pottery motifs encountered at the site. The pot is in the collection of the Illinois State Museum.

Not far from this group burial were the remains of three men and three women; buried with them was a wealth of grave goods [*see illustration on page 100*]. A sheet of rolled-up copper, roughly three feet long and two feet wide, had probably come from the Lake Superior region. Several bushels of sheet mica may have been imported from as far to the east as North Carolina. Quantities of arrowheads were present. Some had been freshly flaked from a variety of local materials. Others, including arrowheads made of a black chert found in Arkansas and Oklahoma, had evidently been imported ready-made. None of the arrowheads showed any sign of ever having been used. Finally, there were 15 beautifully polished double-concave stone disks of the kind European visitors later saw used in a sporting event that the Indians of that time called "chunky."

These rich burials were covered by a mound that stood to the southeast of the original (and now twice extended) platform mound. They constitute the fifth episode of burial at Mound 72. It seems reasonable to suppose this episode was contemporaneous with the sacrifice of the young women.

The sixth episode of burial is less spectacular. Ten oblong pits were dug in an area that was then probably near the southwestern edge of the multiple mound. The pits were used for mass burials and in most instances the individuals were buried soon after death. When the last oblong grave was filled, the builders of Mound 72 covered the multiple structure with fresh earth, giving it the ridgetop form that had initially attracted our attention. From the first episode to the last the sequence of burials and mound building seems to have occupied less than 100 years out of the 150-year Fairmount phase.

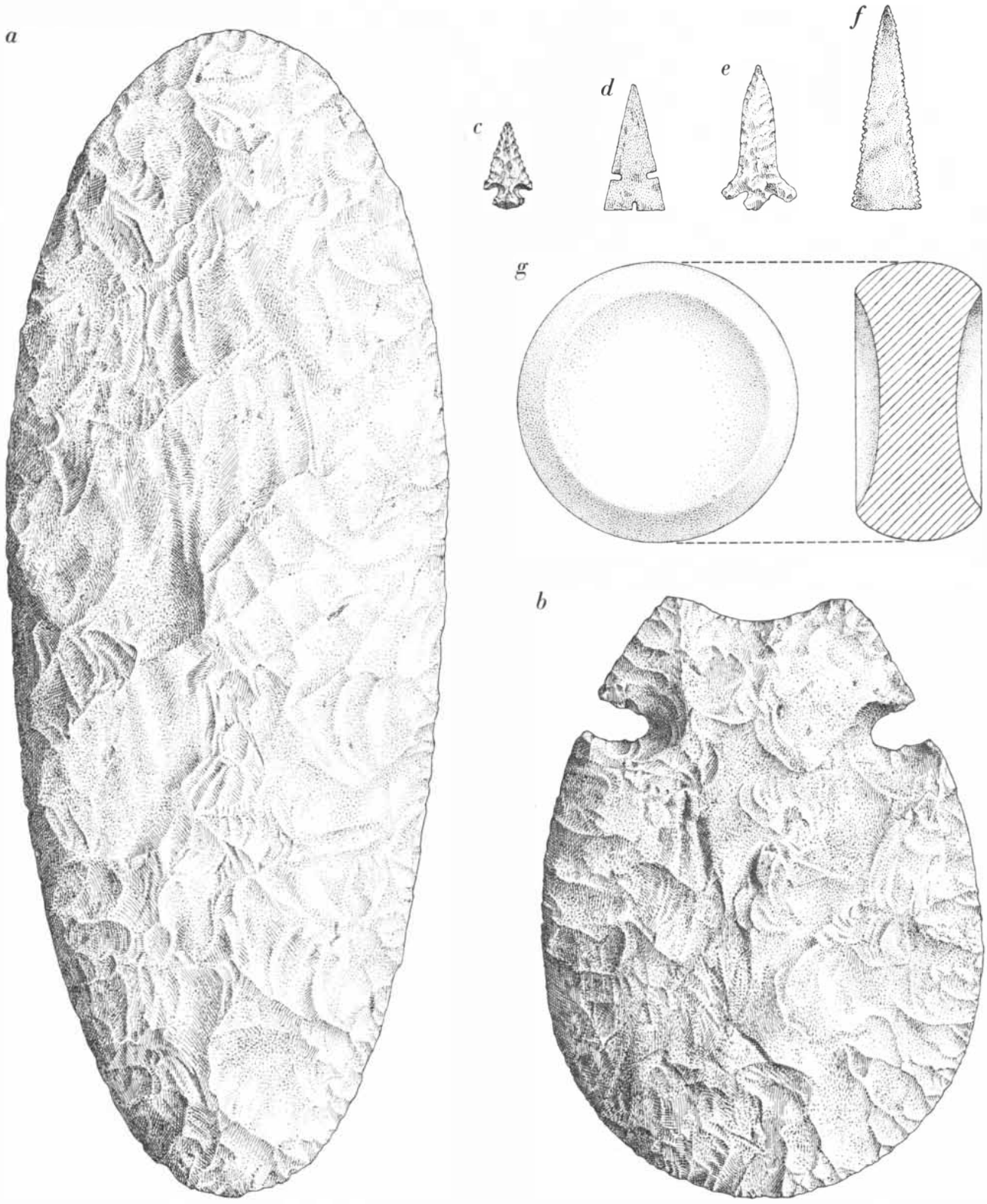
About A.D. 1050 the Fairmount phase was succeeded by a 100-year phase that has been named Stirling. Cahokia continued to grow during this phase, until the settlement covered between four and five square miles. Excavations of the fourth terrace at Monks Mound by groups from Washington University indicate that about A.D. 1100 the terrace was walled and was the site of one large building and several lesser ones. Work by our group has revealed that other buildings stood on the southwest corner of the lowest terrace at that time.

Elsewhere in American Bottoms other centers marked by mounds, in particular the Lunsford-Pulcher site to the south, began to assume increased stature. The mounds of St. Louis and East St. Louis appear to have been started during the Stirling phase, and another multiple-mound community, the Mitchell site, grew up some eight miles north of Cahokia. All four communities were located close to the major waterways of the region. This suggests that waterborne commerce of the kind indicated by the exotic Fairmount-phase grave goods may by then have become as important to the prosperity of American Bottoms as the high-yield harvests from the bottomland farms.

At some time about A.D. 1150 the Stirling phase at Cahokia was succeeded by the Moorehead phase, named after Warren K. Moorehead of the University of Illinois, who studied the Cahokia mounds almost half a century ago. The Cahokia community attained its highest development during that 100-year phase. At Monks Mound a succession of platform mounds, topped by timber buildings, were built on the lowest terrace. Construction activity is also evident elsewhere. One Moorehead-phase project was the rebuilding of an elaborate timber stockade, probably defensive in nature, that at least partly surrounded both Monks Mound and 16 other mounds in its vicinity.

Traces of this structure were first detected in aerial photographs; they appear as white lines running in a north-south direction to the east of Monks Mound. Our group conducted test excavations of the traces in the latter 1960's and discovered evidence of at least four consecutively constructed stockade systems. The first one, apparently built during the Stirling phase, was a wall made of large logs set on end. At intervals the wall incorporated circular projections, or bastions. The second stockade, probably built early in the Moorehead phase, was similar in design except that the bastions, located at the same points as those of the first stockade, were square rather than circular. A third stockade differed from the second only in minor details, and a fourth phase of construction involved an extensive remodeling of the third stockade.

It proved possible to trace the north-south line of the stockade for about 700 yards in the aerial photographs before it turned toward the southwest. Photointerpretation suggests that the stockade, after turning, continued to a point some 300 yards due north of Mound 72 and



STONE ARTIFACTS from Cahokia, mostly chipped out of flint, range in size from large farm tools to small points. All are shown here at 60 percent actual size. The narrow cutting tool (*a*) is more than a foot long and may have been used like a hoe. The notched oval (*b*) more closely resembles a conventional hoe blade; it is some eight inches long. The smallest point (*c*) is one of several found in Mound 72 that were made from a black chert that is not locally available; the points were probably made in Arkansas or Oklahoma and imported to Cahokia. This example is 1.2 inches long. The triply notched point (*d*), 1.5 inches long, is also from Mound 72, as is the peculiarly tanged point (*e*), 1.7 inches long, that may be an import from Arkansas. The largest point (*f*), 2.5 inches long, is also from Mound 72; it is made from a kind of chert found in southern Illinois. The polished stone that resembles a concave hockey puck (*g*) is identical with those used in historical times by the Indians of the Southeast in the game of "chunky." The specimen, 3.5 inches in diameter, is in the Illinois State Museum. The large tools are in the St. Louis Museum of Science and Natural History. The points, unearthed during the author's excavation of Mound 72, are now at the University of Wisconsin at Milwaukee.

then angled back along a northwest leg, roughly equal to the southwest leg in length, before turning due north again. So far nothing is known about the stockade's northern terminus. If its northern limit was located exactly where the northernmost traces of its east side now seem to end, it would have enclosed some 200 acres in the central part of Cahokia.

The stockade has been called a fortification because it incorporated bastions or perhaps towers. Another interpretation of the structure is possible. Monks Mound and 16 other earthworks, including some of the largest platform mounds at the site, lie within the area more or less enclosed by the stockade. Did the structure perhaps screen off and isolate a central core of the community that had a higher status than the periphery? In the absence of the kinds of data that only further excavation can provide, such questions remain unanswered. In any event evidence other than maintenance of the stockade supports the view that the Moorehead phase saw Mississippian culture reach its peak at American Bottoms. For example, the woodhenge excavations suggest that the last of the solar observatories was built at this time. Again, to judge by the number of house sites in areas where such sites have been found, the Moorehead-phase population of Cahokia may have approached 40,000.

How was this substantial population distributed? The area outside the stockade, with its many mounds, suggests a pattern. The mounds appear to have been organized in clusters, and each cluster includes platform mounds, plazas and what are probably burial mounds. The clusters suggest the existence of subcommunities located within the larger metropolis.

The distribution of house sites at Cahokia is also suggestive of a pattern of community organization. A ridge runs through the site from east to west; it is along this ridge line, over an area of some 2,000 acres, that the main concentration of housing is found. The houses were spaced at regular intervals, several to an acre. Of pole-and-thatch construction, they were mostly rectangular in floor plan. Evidently once a building site was chosen a succession of structures were built in the same place, suggesting several generations of occupation. The houses show substantial variations in size. Some could have been the residences of persons of high status, while others may have sheltered craftsmen or even farmers.

Finally, evidence of change in the pattern of land utilization, particularly close to the stockade, hints at the social complexity of Cahokia at its height. In one location some 400 yards west of Monks Mound land that was residential at an earlier time was transformed during the Stirling phase into an area of walled enclosures and large public structures. The area was not returned to residential use until after A.D. 1250. The construction of the stockade provides another example. A part of the timber wall appears to have been built through the middle of an active residential area without regard for the residents. All these findings hint at the power of the central authority that directed the destiny of the Cahokia community.

RICH CACHE OF GRAVE GOODS was found at Mound 72 together with the bones of six individuals. Above the marker arrow (*lower left*) are a cluster of arrowheads, including many imported from the Ozarks. Above them (*center*) is a group of chunky stones. The cylindrical mass to their right is a rolled sheet of copper from the region of Lake Superior. Parallel with the roll of copper is a row of large and small shell beads; they are from the Gulf coast. The rounded pile just above the chunky stones is a mass of imported sheet mica.

The next phase at American Bottoms, named Sand Prairie, continued from A.D. 1250 until about 1500. The data from Cahokia for this period give the impression of a far lower level of activity. It was now that the area to the west of Monks Mound was reconverted from public uses to private ones and that minor additions were made to the lower part of Monks Mound. At least one mound still supported a public building. This is the Murdock mound, a double-platform structure located at the angle where the east wall of the stockade turns to the southwest. Excavation there by workers from the Illinois State Museum in 1941 uncovered timbers from what was apparently the last stage of construction of a large building. Wood samples gave a carbon-14 age ranging from A.D. 1270 to 1470.

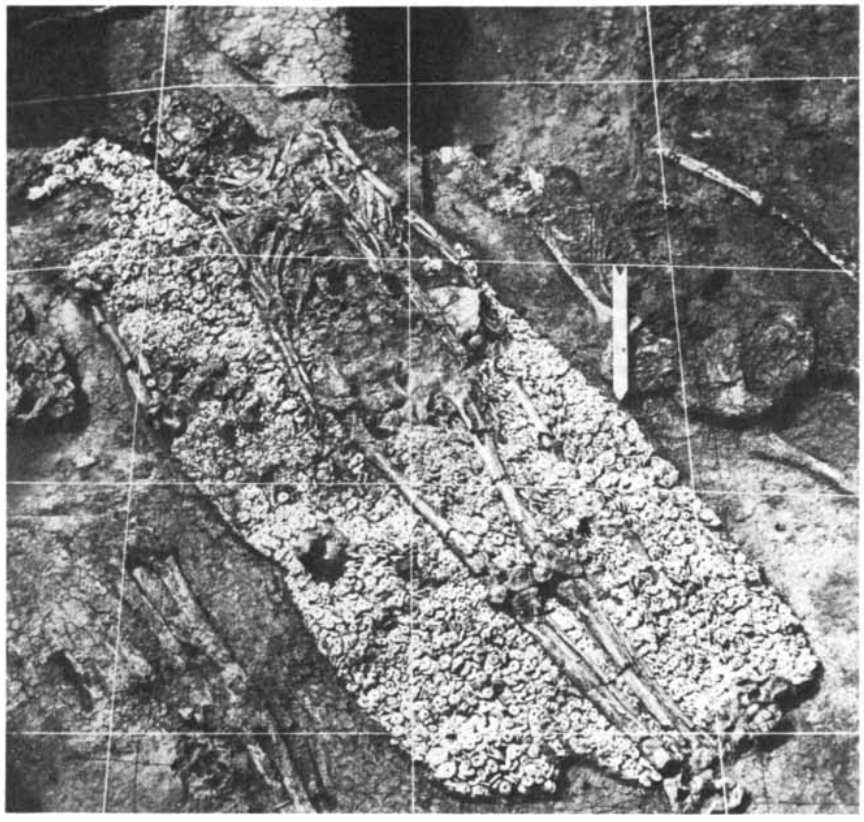
No name has yet been chosen for the final phase at American Bottoms, which extends from A.D. 1500 to 1700, or about the time of contact with Europeans. The pattern, insofar as it has been traced, is one of continuing decline. About all that is certain, as excavations have shown, is that local Indians sometimes visited Monks Mound at the beginning of the 18th century to bury their dead.

Why did the remarkably successful Mississippian culture at American Bottoms fade away? Perhaps the decline was related to the exhaustion of local resources: a lack of timber for public and private buildings, for the stockade and for the sun circle, and the disappearance of the game animals (mainly deer) in the immediate hinterland that had provided a vital part of the inhabi-

tants' diet. Even the fertility of the bottomland loam, where the maize was grown, was not inexhaustible. Evidence from elsewhere in North America, for example at Chaco Canyon in New Mexico, indicates that population concentrations and the overuse of local resources go hand in hand.

It is also possible that the decline of the large settlements at American Bottoms, Cahokia in particular, was related to the growing strength of other centers of Mississippian culture. Cahokia had been the largest single Mississippian center and also perhaps the earliest. Its role in the evolution of the Mississippian way of life must have been dominant for centuries. At the same time American Bottoms was a major crossroads, particularly for travel by water. Its northern extremity was marked by the confluence of the Illinois and Mississippi rivers; the Illinois was a highway to the north-northeast and the upper Mississippi was a highway to the north and north-northwest. Just to the south of this confluence the Missouri River, a highway to the west and northwest, also joins the Mississippi. Toward the southern end of American Bottoms a highway into the Ozarks, the Meramec River, enters the Mississippi from the west. And some 150 miles to the south is the confluence of the Mississippi with the second-greatest of North American rivers, the Ohio, a major highway leading to the east and northeast.

The exotic materials found at sites in American Bottoms (for example black chert, probably from the Ozarks; native copper, almost certainly from Lake Superior; sheet mica, possibly from North Carolina; salt from southern Illinois or Missouri; lead from northern Illinois, and marine shells from the Gulf coast) are concrete evidence of the waterborne commerce that presumably kept American Bottoms in contact with other Mississippian or proto-Mississippian areas of North America. In addition many perishable goods, of which no archaeological trace remains, probably moved through Cahokia in the course of this commerce. Meanwhile Mississippian culture expanded to other areas and thrived. When the first Europeans traveled through southeastern North America, even though Cahokia was by then abandoned and unknown, they found many flourishing Mississippian regional centers. The young warriors were playing games with chunky stones, their elders were tending charnel houses and preparing bundles of bones for final burial and (among the Natchez Indians) an elaborate social hierarchy, headed by chiefs called suns, still built mounds and lived on top of them and regularly offered human sacrifices in the ceremonies that attended the burial of people of high status.



BURIAL PLATFORM made of thousands of shell beads, drilled for stringing, supports the skeleton of a man who was buried in the extended position. The burial was found in the part of Mound 72 near where a large post had once stood to mark the north-south axis of Cahokia. The interred man probably held an important social position. All the burials in and under Mound 72 date from the Fairmount phase at Cahokia: A.D. 900 to A.D. 1050.

It is plausible to suppose that in addition to the handicap of diminishing resources the Mississippians of American Bottoms suffered from a loss of social status and economic power as the other regional centers of Mississippian culture developed their own hinterlands and spheres of influence. This does not alter the fact that the settlements at American Bottoms and at Cahokia in particular offer a nearly unique opportunity to study the rise of a complex society that for centuries controlled the natural resources of an immediate hinterland and also oversaw the distribution of highly valued resources drawn from more distant areas. The processes that led to the rise of such a society, although they are not yet understood in detail, can be suggested in broad outline.

The initial stimulus seems to have been the population expansion and the jump in agricultural productivity. High crop yields and sedentary communities in turn combined to foster further increases in population density; perhaps it was competition for available farmland and the resulting conflict between rival communities that stimulated the evolution of social controls and societal hierarchies.

The fact that one community in American Bottoms, Cahokia, became the dominant community can be understood in these terms. First, Cahokia stands on some of the best agricultural land in the region. Second, a network of sloughs, lakes and creeks gave Cahokia easy access both to the rest of American Bottoms and to the long-distance transport network formed by the big rivers. Cahokia was thus a central place ideally situated both to exploit the resources of the immediate hinterland and to dominate the trade in exotic goods.

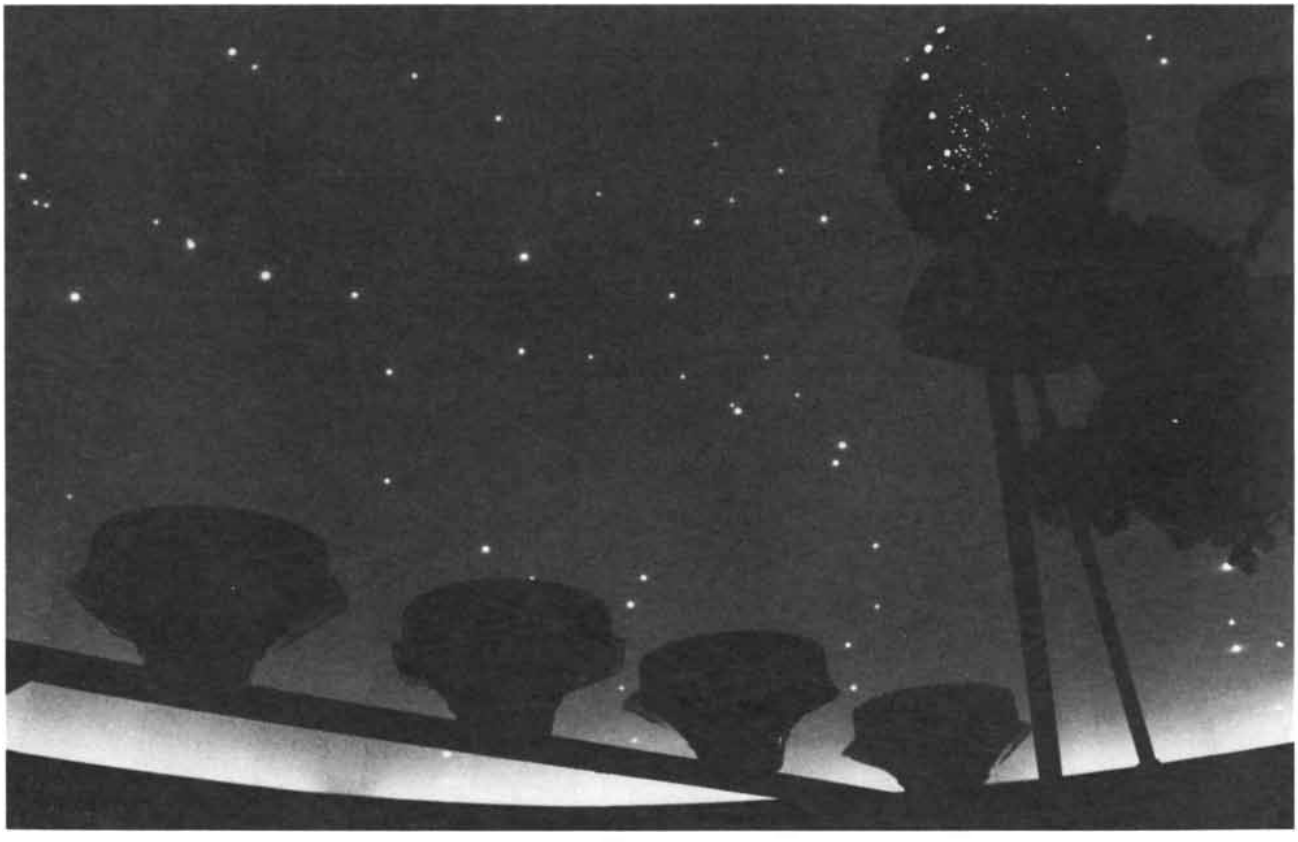
So much for the rise of Cahokia, at least in the light of present knowledge. The processes involved in its decline, however, need to be understood quite as much as the processes involved in its rise. Those who plan future studies of American Bottoms should recognize that the question of why Cahokia was abandoned is among the most significant questions that remain unanswered.

The Stellar-Orientation System of a Migratory Bird

When the indigo bunting is put in a planetarium, it exhibits an ability to orient itself by the stars. This, however, can be only one of the cues it uses for long-distance navigation

by Stephen T. Emlen

The blackpoll warbler is a small, inconspicuous songbird that breeds during the summer in the stunted conifer forests of Alaska and northern Canada. When fall approaches, these birds embark on a remarkable migratory journey. They leave the northern forests and fly east-southeast across the North American continent to the Atlantic coast. During this stage of the journey they stop to feed and build up stores of subcutaneous fat that will serve as a vital energy reserve for the next stage. The blackpolls concentrate near the coast of New England and the Maritime Provinces of Canada, waiting for the right weather conditions. Then, as the next high-pressure cell moves in from the west, bringing with it winds from the north or northwest, the blackpolls depart again, this time over the ocean on a nonstop flight that will take them three to five days. They fly over Bermuda, the Antilles and Puerto Rico, stopping only when they make landfall on the northeastern coast of South America. It is a tremendous feat: a nonstop flight of



EXPERIMENTAL ARRANGEMENT in a planetarium for testing the ability of the indigo bunting to orient by the stars is shown. The photograph was made in the Southern Cayuga Atmospherium-Planetarium in Poplar Ridge, N.Y. Projectors are not visible to birds in the cages. Stars at any latitude and longitude, as well as celestial motion, can be projected onto the planetarium dome.

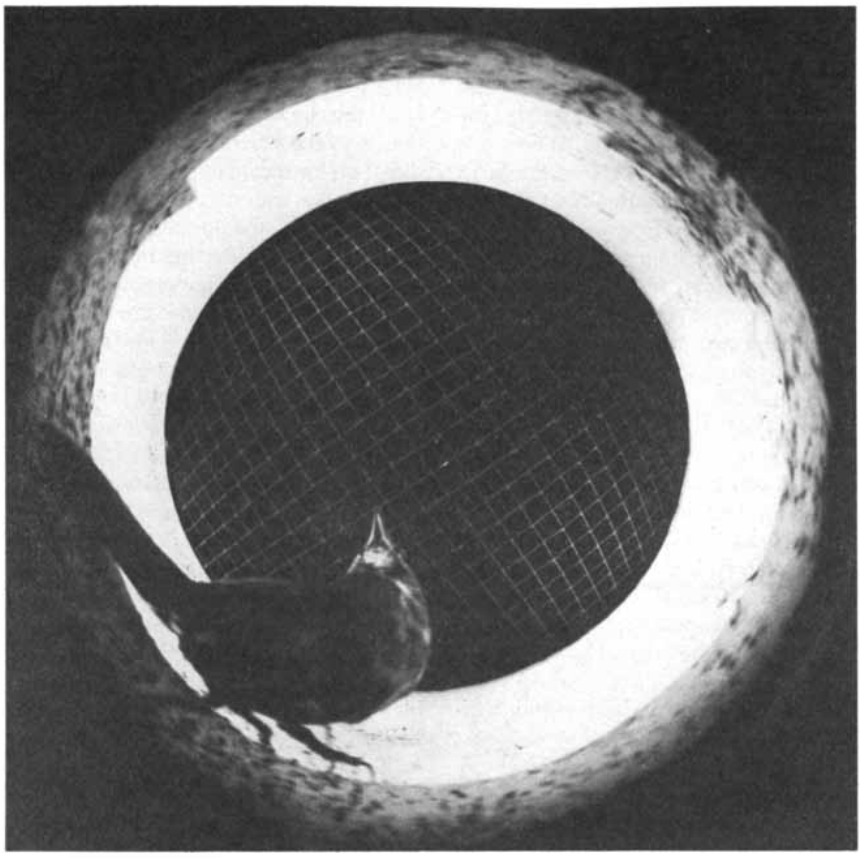
more than 2,400 miles over water by a bird weighing less than 20 grams! Any error in navigation would obviously lead to disastrous results. Any misinterpretation of the weather would also lead to disaster, since a blackpoll that encounters stiff headwinds or a storm has no chance of landing to rest or to find shelter.

The blackpoll warbler is an extreme but not atypical case of birds that migrate south. Fully two-thirds of the species of songbirds that breed in the northern U.S. travel south in the winter. The distances of migration typically range from 600 to 1,800 miles, but each fall some songbirds make one-way trips of up to 4,000 miles. The following spring the birds fly back to their breeding grounds. Year after year the adult birds return with amazing precision to the same several square miles of territory at both their breeding and their wintering grounds.

How do the migrating birds select the appropriate flight directions? Can they determine when they have been blown off course, and can they correct appropriately? How do they know when they have arrived at the latitude of their destination? What is the fate of young birds flying alone on their first migratory trip?

Biologists have long been intrigued by the phenomenon of bird migration, but it is only during the past two decades that significant progress has been made toward answering the fundamental questions about it. Today scientists around the world are focusing their attention on questions of animal navigation. Hundreds of thousands of migrant birds are being individually marked with a leg band so that field investigators can determine their migratory paths by plotting their recapture locations. In the laboratory other workers are testing the ability of birds to detect different potential directional cues and are examining how such cues are used. Ornithologists are tracking "unseen" migrating birds with radar and following individual migrants by attaching small radio transmitters to them.

The results of these various studies have shown that bird navigation is not a simple affair; it is not entirely dependent on any single cue or sensory system. It seems that migrating birds make use of a variety of cues to determine their direction and maintain it in flight. A recent article by my colleague William T. Keeton describes some of the types of cues available and the interplay among them [see "The Mystery of Pigeon Homing," by William T. Keeton; SCIENTIFIC AMERICAN, December, 1974]. Here I shall concentrate on one cue system that has been studied intensively and that appears to be of major importance to night-migrating birds: orientation by the stars. The reader should bear in mind that I am covering only one aspect of how birds navigate. It is only by dissecting out the various aspects and studying them one at a time that an understanding of the full story will ultimately be achieved.



INDIGO BUNTING is shown hopping onto the side of a test cage. The photograph was made by placing a camera with a "fish-eye" lens at the bottom of the cage. Black marks on the white side of the cage are footprints made by the bird. The screen is the top of the cage.

The scientific study of the directional orientation of migratory birds had its breakthrough with the pioneering work of Gustav Kramer, an ornithologist at the Max Planck Institute for Marine Biology at Wilhelmshaven. It had long been known that when migrant songbirds were kept in captivity, they displayed intense activity at night during the periods of their normal spring and fall migration. When Kramer placed songbirds in circular cages during their migratory period, he discovered that they would spontaneously orient their activity in a particular direction. By manipulation of the cues available to the caged birds the determinants of direction finding could be studied. Through this technique migration could be "brought into the laboratory."

In the late 1950's another German ornithologist, E. G. F. Sauer of the University of Freiburg, carried out a long series of experiments with European warblers. It was he who first hypothesized that these warblers determine their migratory direction from the stars in the night sky. Since then much has been learned about star orientation by birds. Numerous species have been examined, and it appears that the ability to orient by the stars is widespread among birds that migrate at night.

For some years I have been studying one night-flying migratory bird, the North American indigo bunting (*Passerina cyanea*). Enough information is now available to piece together a fairly complete picture of how the stellar-orientation system in this species operates. The species breeds throughout the eastern U.S., where the brilliant blue male is a well-known songster. During the fall migration the buntings fly up to 2,000 miles to winter in the Bahamas and in southern Mexico and Central America south to Panama.

When indigo buntings are kept in captivity and exposed to the same pat-

tern of day lengths they would experience in nature, they exhibit intense nocturnal activity in April and May and again in September and October, the times of normal migration. When this restlessness appeared, I tested the birds by placing them individually in circular cages. The cage was constructed of a piece of white blotting paper, rolled and stapled to form a funnel, mounted on a base consisting of an ink pad and topped with either a clear plastic sheet or a hardware cloth screen. A bird inside the cage can see only the sky overhead, since all ground objects are blocked from view.

A bunting in migratory condition stands in one place or turns slowly in a circle, its bill tilted upward and its wings partly spread and quivering rapidly. At frequent intervals the bird hops onto the sloping paper funnel, only to slide back and continue its pointing and quivering. Each hop from the ink pad leaves a black print on the paper. The accumulation of inked footprints provides a simple record of the bird's activity; they can later be counted and analyzed statistically.

In the first stage of my studies I placed the buntings in their funnel cages outdoors on clear, moonless nights. In September and October most of the birds exhibited a distinct preference for jumping toward the southern sector of the cage, the same direction in which they would have migrated if they could. When the birds are tested in late April and May, however, the preferred direction is to the northeast, the appropriate direction for spring migration.

Since the wall of the test cage completely screens the horizon from view, it seems likely that the buntings are able to determine their migration direction when the only visual cues are those provided by the night sky. This hypothesis was reinforced by the changes observed when the birds were placed outdoors at night under cloudy conditions. As the stars disappeared behind the clouds, the orientation of the birds deteriorated considerably [*see lower illustration on opposite page*].

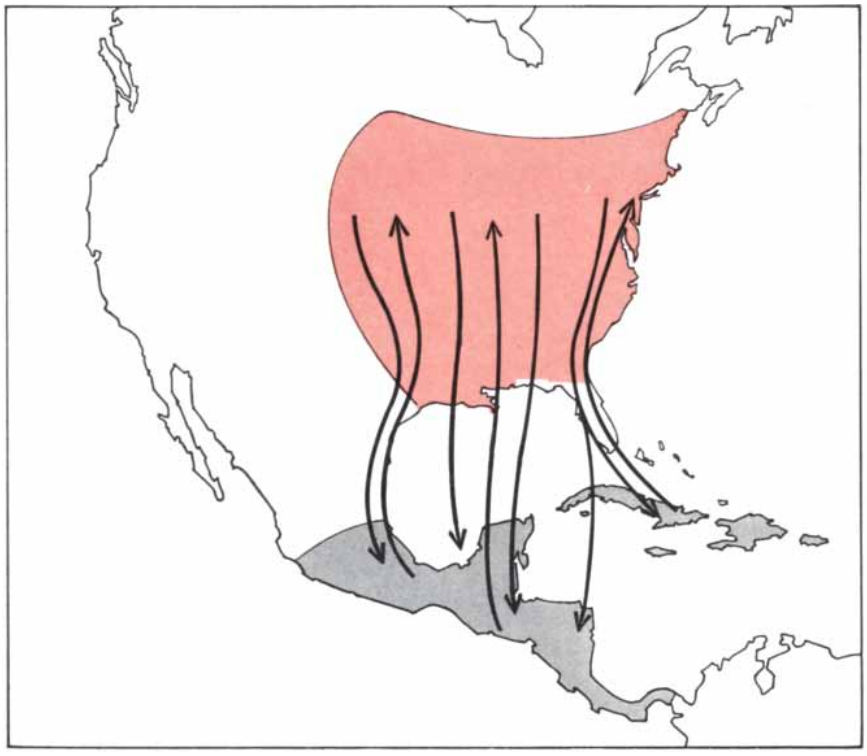
In order to test the stellar hypothesis under more rigorous conditions, I took the buntings into a planetarium, an approach also used by Sauer in some of his experiments with European warblers. In September and October, when the birds were exhibiting nocturnal restlessness, I projected the normal fall stars onto the planetarium dome. When the buntings were tested in the funnel cages, they oriented to the south. Birds tested in April and May under a spring sky in the planetarium consistently oriented to the north and northeast. When the North Star in the artificial sky was shifted to the east or west, the buntings changed their orientation to the new "south" or the new "north," depending on the season of the year. The change in orientation behavior was consistent and predictable. In control experiments I turned off the star projector in the planetarium and exposed the birds to a diffusely illuminated dome. Their behavior paralleled their response to overcast conditions outdoors: the accuracy of their orientation deteriorated considerably.

Since indigo buntings are willing to exhibit meaningful orientation in spite of the confinement of captivity, I was able to further modify the visual cues in the night sky and thus dissect out the detailed workings of the birds' stellar-orientation system. The experiments were designed to answer the following questions: Which stellar cues are important? How are such cues employed? What kind of information does the bird obtain? How accurately does it obtain it?

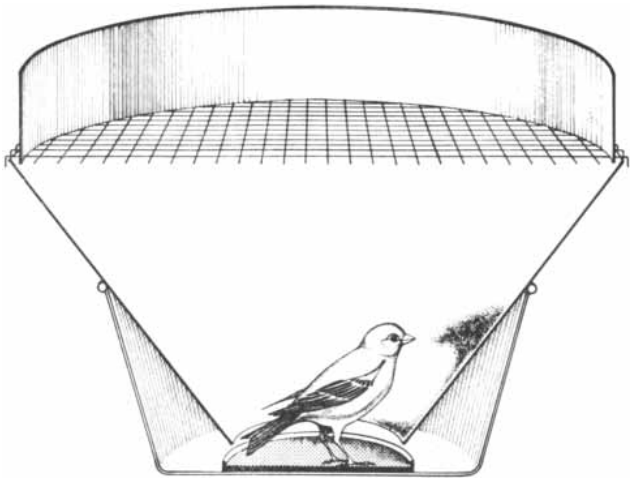
In theory there are two ways birds could determine direction by the stars. One way would be for the bird to locate a critical star or a group of stars and then guide itself by flying at a particular angle with respect to the star or the group of stars. The absolute position of a star, however, is not constant throughout the night: stars shift from east to west as the result of the rotation of the earth. In order to maintain a given compass direction the bird would have to alter its angle of flight with respect to the selected star in such a way as to compensate for the apparent motion of the star [*see top illustration on page 107*]. Such a mechanism would be analogous to the sun-compass orientation in which a daytime bird migrant, making use of an internal time sense, correctly compensates for the daily movement of the sun across the sky.

The requirements for a stellar-navigation system are much more demanding than those for a system that depends on the sun for determining compass direction by day. There is only one sun, and it moves at a regular rate, but there are thousands of stars and different stars are visible above the horizon at different times of the night and at different seasons. A nocturnal migrant presumably would need to be able to consistently locate a specific star or a specific group of stars, and that would require it to possess some form of pattern recognition.

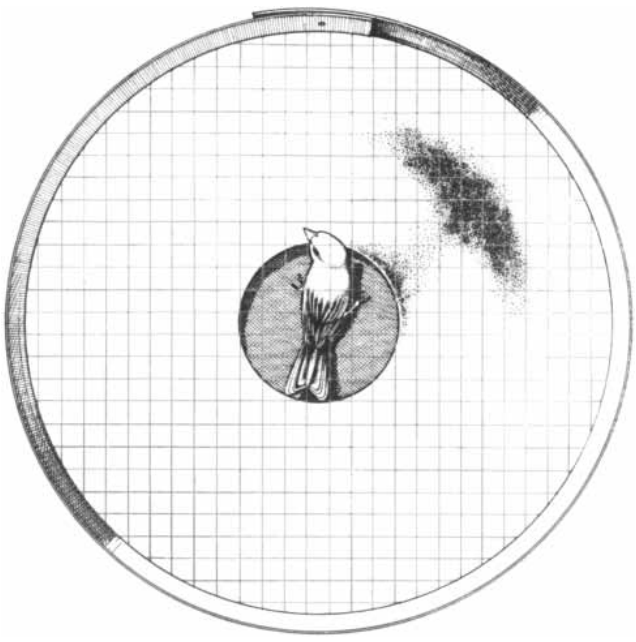
In addition the rate of compensation for apparent motion will differ, depending on the star or stars selected. Celestial motion is an apparent motion produced



MIGRATION OF INDIGO BUNTINGS proceeds along a broad front. Buntings migrate from their wintering grounds in the Bahamas, southern Mexico and Central America (*gray areas*) in late April and arrive at their breeding grounds in the eastern U.S. (*colored area*) throughout the month of May. They depart for wintering grounds in September and October.



CIRCULAR TEST CAGE for determining the directional preference of an indigo bunting is shown in cross section and in a top view. Funnel portion of the cage is made of white blotting paper.



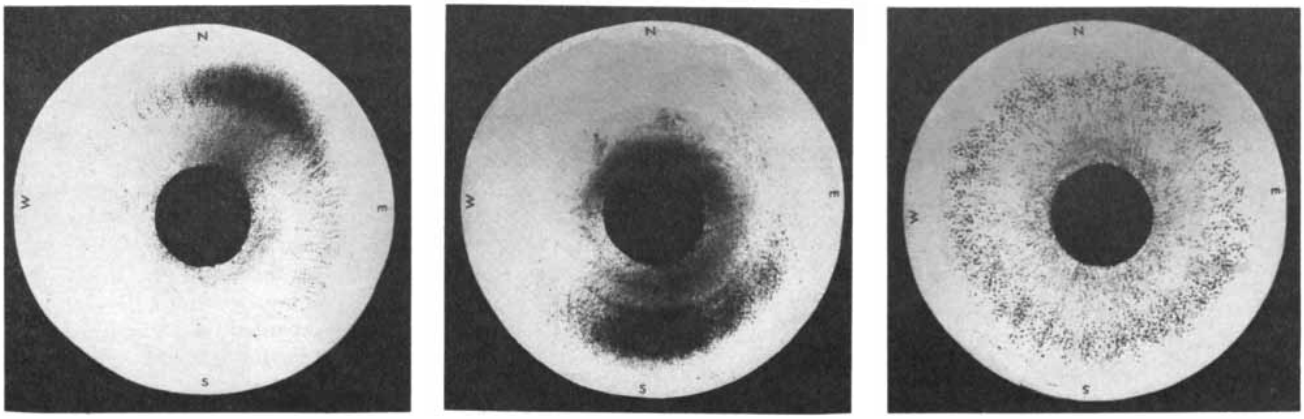
The bunting stands on an ink pad, and each time it hops onto the sloping funnel wall it leaves black footprints. The bird's view is limited to a 140-degree overhead sector of sky when it hops up.

by the earth's rotation once every 24 hours. All stars move with an angular velocity of 15 degrees per hour, as the sun does. The linear velocity of a star will vary, however, depending on how close it is to the North Star. Stars near the North Star move through a small arc, whereas stars near the celestial equator move through a large one. If a bird were to use star groups in different parts of the sky, it would have to compensate at a different rate for each group. Finally, the direction of compensation depends on whether the guiding stars are in the north or in the south: northern stars would require clockwise compensation, southern stars counterclockwise.

According to the second theoretical model, the bird would use patterns of stars to determine directional reference points. Human beings easily recognize the Big Dipper by the characteristic arrangement of its stars. And by visually extending the line joining the two pointer stars of the Big Dipper, they can readily locate the North Star and hence geographic north. Star patterns such as the Big Dipper also move across the sky, but the shape of each pattern remains constant and each preserves a distinct relation to the North Star.

Since each star bears a fixed geometric relation to other stars, it would be theoretically possible for a bird to determine a directional reference point from any number of star patterns [*see middle illustration on page 107*]. The major difference between this model and the first one is that a compass direction can be determined from the geometric patterns of the stars independently of an internal time sense, or "biological clock."

I tested these alternative hypotheses in a planetarium by creating an artificial situation in which the astronomical time would be out of phase with a bird's internal time sense. If a time sense is involved in star navigation, then presenting the stars in positions that are advanced or retarded from local time should cause the bird to orient in the wrong direction. On the other hand, there should be no change if the bird relies only on the geometric patterns of the



FOOTPRINT RECORDS of male indigo buntings tested in a circular cage placed outdoors under the stars on moonless nights are shown. In the spring the bird typically orients its hopping to the north (*left*). In the fall its hopping is oriented to the south (*middle*). When the stars are obscured by clouds, the bunting's activity remains high but orientation of the hopping is random (*right*).

stars. I tested buntings when the planetarium sky was three, six and 12 hours ahead of local time and when it was three, six and 12 hours behind local time. The results were clear: the birds generally maintained their normal migratory orientation under all these conditions. Apparently indigo buntings do not make use of an internal time sense to orient by the stars but obtain directional information from star patterns, much as human beings do.

I then turned to the question of which star groupings are of particular importance to buntings. Once again the planetarium was an invaluable tool, since one can block from view selected stars, constellations or entire areas of the sky. In a series of experiments I systematically removed and later reinserted portions of the artificial sky. I found that most buntings rely for direction finding on the northern area of the sky that lies within about 35 degrees of the North Star. The major constellations in this area are the Big Dipper, the Little Dipper, Draco, Cepheus and Cassiopeia. The birds relied on the northern circumpolar stars not only during the spring migrating season when they normally fly north but also during the fall migrating season when they fly south.

An important corollary finding was that there is considerable redundancy in the buntings' recognition of star patterns. The birds are familiar with several star groups, and the removal of one group of stars, say the Big Dipper, often merely forces them to rely on some alternate constellation. Since birds frequently migrate on nights when there is variable cloud cover, such redundancy is obviously adaptive.

Navigation can be regarded as involving a two-step process. Consider a man equipped with a map and a compass. In order to determine how to reach a particular geographic destination he must first calculate his position on the map with respect to his goal and then use the compass to select the appropriate direction. The navigation problems of a migrating bird can be viewed in the same way, that is, as a "map and compass" process.

Theoretically an accurate knowledge of the absolute positions of the stars coupled with a stable and highly accurate internal time sense could provide enough map-and-compass information for a bird to determine its absolute geographic position. If the bird retains a precise memory of the temporal position of the stars at its destination point, it could in principle select the appropriate direction to the goal from the displacement of the stars in the sky overhead.

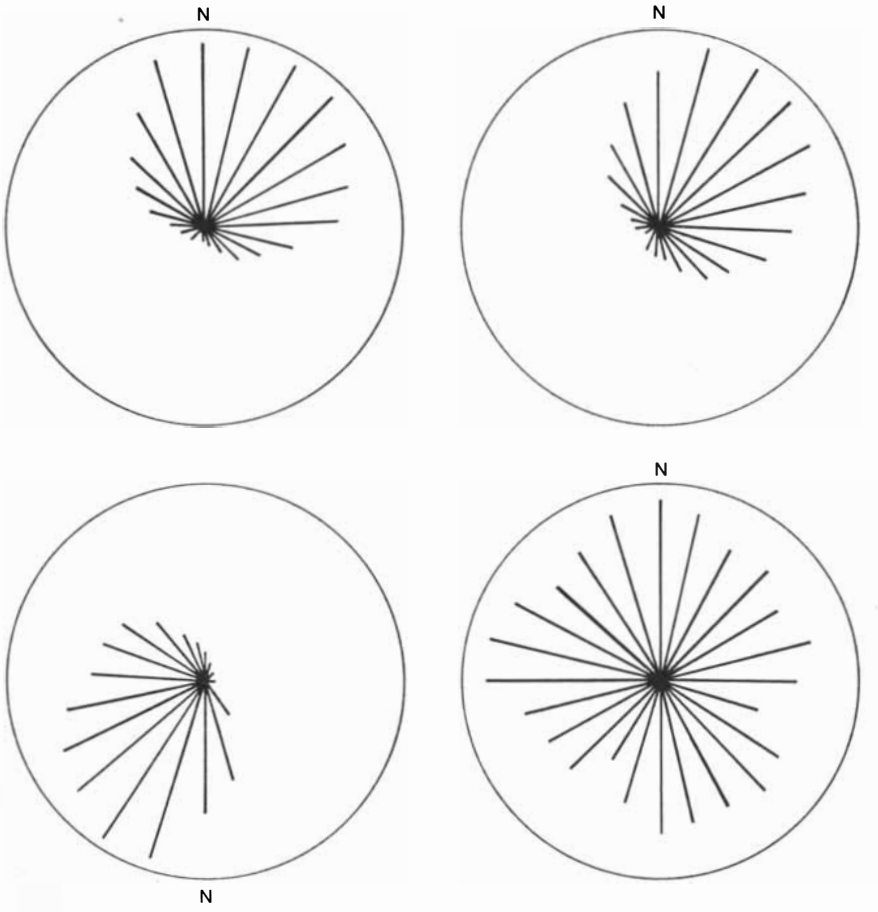
The finding that the indigo bunting does not integrate a temporal component with its use of stellar cues implies, however, that the bird does not detect or correct for longitudinal displacement, at least not by celestial cues. Hence star orientation in the indigo bunting appears to be a compass sense that enables the bird to select and maintain a particular direction but does not provide the information that makes it orient to a particular goal.

What, then, does determine how the star compass will be used? The bunting may be able to locate the Big Dipper or other constellations, but why does it use them to orient north or south rather than east or west? And how does the bunting select north in the spring and south in the fall?

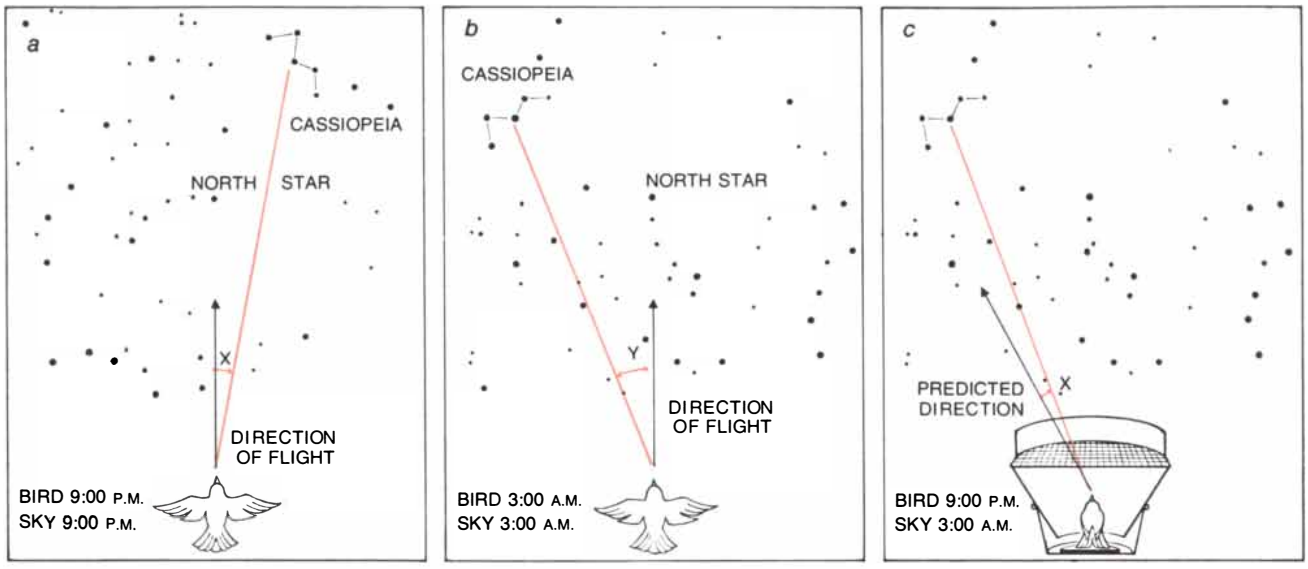
The sidereal, or astronomical, day is four minutes shorter than the solar day. Because of this inequality the temporal positions of the stars change with the seasons, with the result that the stellar information available from the fall night sky is quite different from that available from the spring night sky. Does the indigo bunting have a specific northerly directional response to the stellar stimuli in the spring sky and a southerly response to the stellar stimuli in the fall sky?

To test this possibility I captured 15 adult male buntings in their summer breeding territories near Ithaca, N.Y., and divided them into two groups. The weight, fat level and molt status of each bird was recorded weekly until the testing period the following spring.

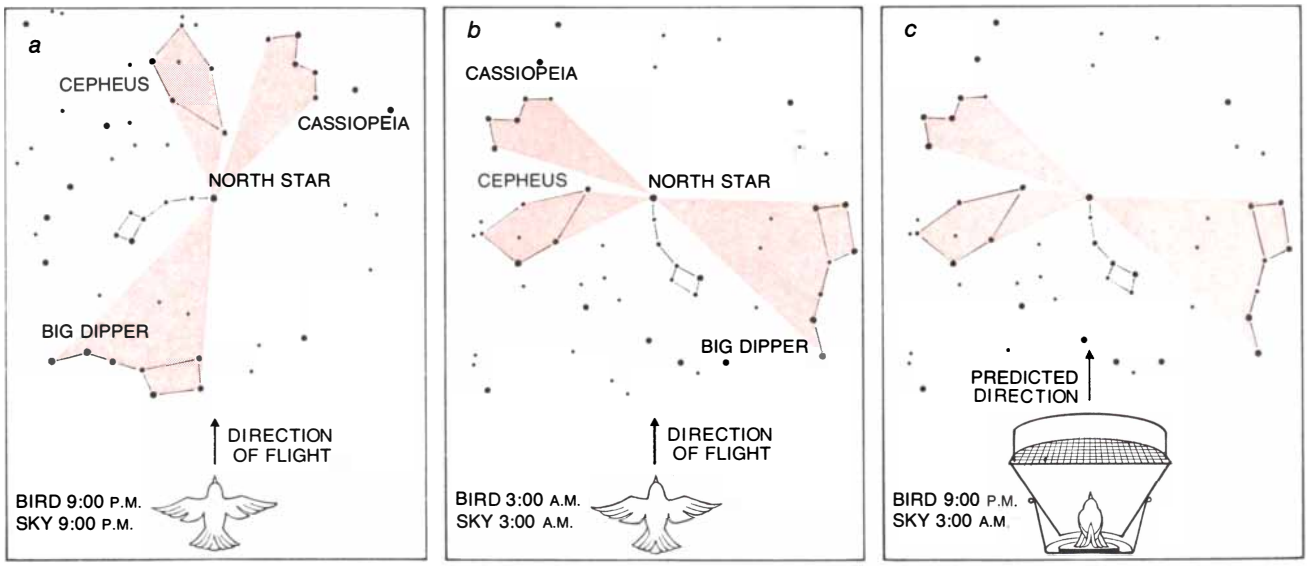
The control group of eight birds lived in a flight room where the length of the day simulated what they would have normally encountered in nature. An astronomical time clock maintained a day length equivalent to the day length at



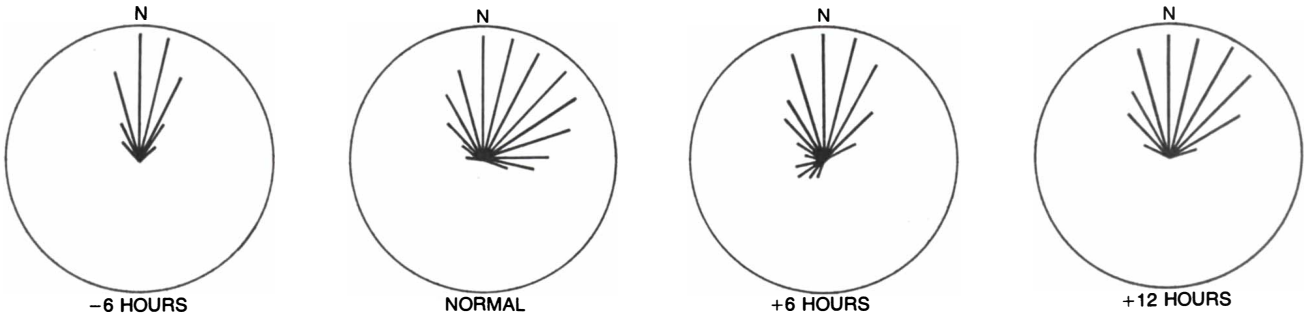
VECTOR DIAGRAMS show the similarity of the orientation of a bunting outdoors under the stars in the spring (*top left*) and under a simulated spring night sky in a planetarium (*top right*). When the planetarium stars were shifted so that the North Star was at the true south, the bird reversed its orientation (*bottom left*). When the stars were turned off and the planetarium was diffusely illuminated, the bunting's orientation became random (*bottom right*). The radius of each circle is equal to the largest amount of activity in any one 15-degree sector, and the vectors for the other 15-degree sectors are proportional to the radius.



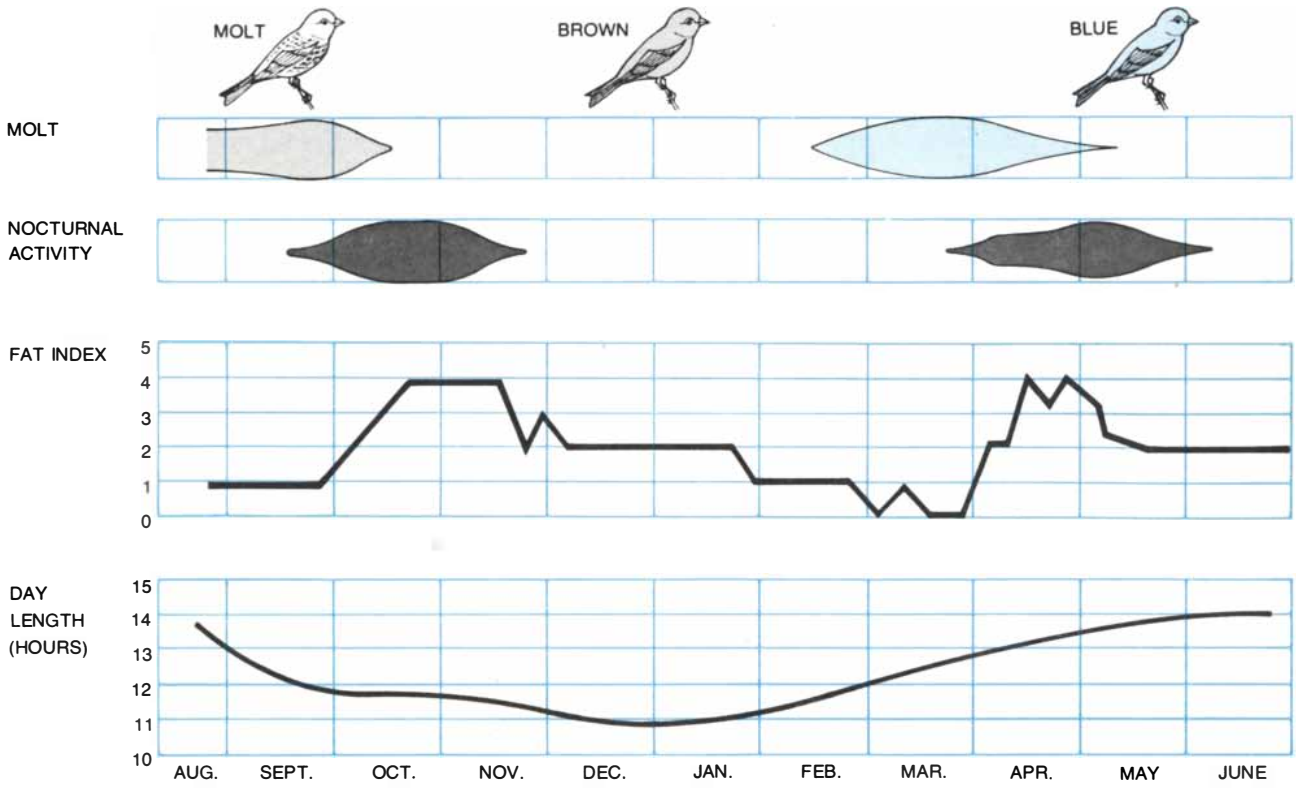
STELLAR-ORIENTATION HYPOTHESIS I proposes that the indigo bunting guides itself by flying at an angle to a particular star or group of stars. Since the positions of the stars change throughout the night, the bird would have to use an internal time sense to compensate for the motion of the stars. For example, a bunting going north at 9:00 P.M. would fly at angle *X* with respect to a critical star (*a*). At 3:00 A.M. the bird compensates for the rotation of the stars by flying at angle *Y* to the critical star (*b*). According to the hypothesis, when a bunting whose physiological time is at 9:00 P.M. is presented with a 3:00 A.M. star pattern in a planetarium, it should compensate in the wrong direction, that is, it should orient at angle *X* with respect to the critical star instead of at angle *Y* (*c*).



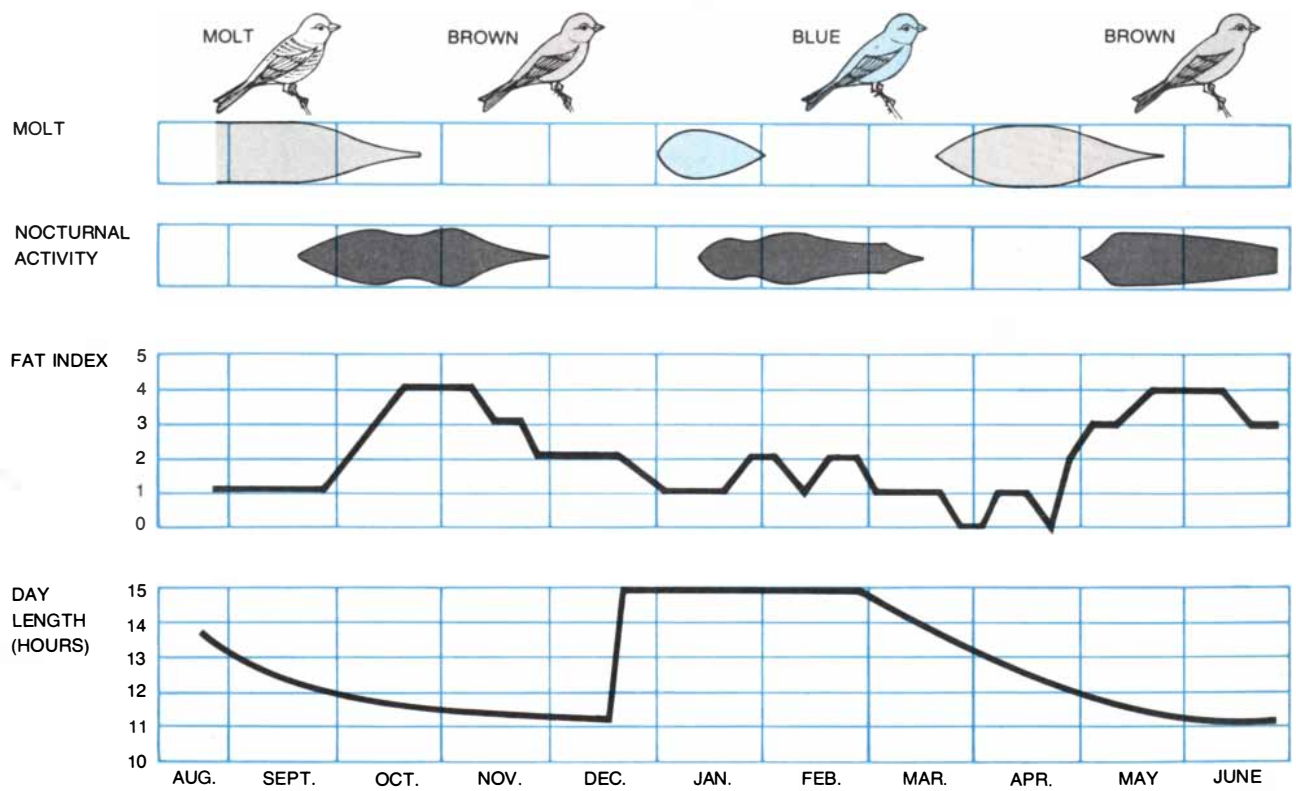
STELLAR-ORIENTATION HYPOTHESIS II states that the bunting obtains directional information from the configuration of the stars. The bird can determine a reference direction such as north from fixed geometric relation of the stars regardless of the time of night (*a*, *b*). When the bunting is exposed to a time-shifted sky in a planetarium, there should be no change in its orientation (*c*).



RESULTS OF PLANETARIUM TEST of the two star-navigation hypotheses show that buntings continue to orient correctly regardless of whether planetarium stars are shifted ahead of or behind the bird's normal physiological time. This indicates that the bunting does not incorporate its biological clock in the star-orientation process and obtains only directional information from star patterns.



CONTROL GROUP OF ADULT MALE BUNTINGS lived in a flight room where an astronomical time clock maintained the day length equivalent to length at their wintering grounds. The birds molted normally in the fall and again in the spring. After each molt they built up fat reserves and became active at night. In May directional preference of the birds was tested in a planetarium.

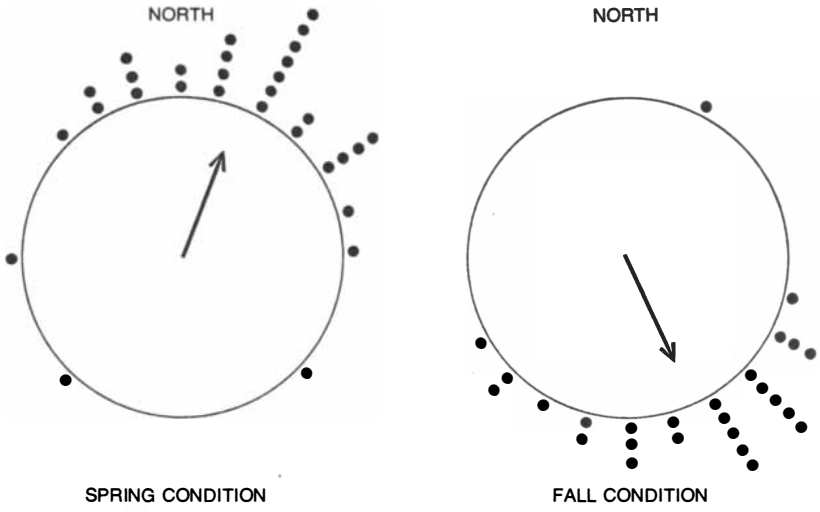


EXPERIMENTAL GROUP OF ADULT BUNTINGS was given an accelerated photoperiod regimen. Beginning in mid-December birds were exposed to the longer day lengths typical of spring. They molted into their blue spring plumage in January. The day lengths were shortened in March, and the birds molted into their brown winter plumage. Their directional orientation was tested in May.

their wintering ground in Guatemala. The male buntings in this group molted normally and acquired their bright blue prenuptial plumage between February and early April. After molting they built up substantial reserves of subcutaneous fat. They became nocturnally restless in April and May.

The experimental group of seven buntings was subjected to the same day-length regime as the control group until mid-December. The birds were then exposed to a spring day length of 15 hours, which caused them to molt into their blue spring plumage in January. Beginning on March 1 the day length was progressively shortened to simulate the day lengths of fall in the buntings' summer breeding territory, and the birds molted out of their blue plumage and into their brown winter coloration. After the molt the buntings built up fat reserves, and nocturnal activity began in May.

The directional preferences of the control group and the experimental group were tested in the planetarium under identical spring night skies. There was a marked dichotomy in the orientation behavior of the two groups. The control group of blue buntings, which were ready for their normal spring migration, oriented to the north and northeast, whereas the experimental group of brown buntings, which were ready for their normal fall migration, oriented to the south.



DIFFERENCE IN ORIENTATION of indigo buntings in the control group (*top illustration on opposite page*) and the experimental group (*bottom illustration*) shows that the physiological state of the bunting affects its migratory direction. The control group, which was physiologically in the normal spring migratory condition, oriented to the north-northeast. The experimental group, which was in an induced fall migratory condition, oriented to the south-southeast. The mean vector of the birds in each group is shown by the arrow. The length of the vector represents degree of agreement among the birds in selecting a direction.

The results indicate that the physiological state of the indigo bunting affects its migratory orientation. We have already seen that the same northern circumpolar stars are used as the chief stellar reference in both spring and fall. It now appears that the polarity of the migratory orientation—whether it is toward or away from the northern circumpolar stars—is under hormonal control. Recent studies by Albert H. Meier and D. D. Martin of Louisiana State University support this hypothesis. They report being able to reverse the orientation behavior of another nocturnal migrant, the white-throated sparrow, by altering its physiological state with the administration of two hormones, prolactin and corticosterone, which appear to have a synergistic effect in stimulating the birds' migratory activity.

The finding that the physiological state of the indigo bunting affects the direction in which it orients does not, however, fully answer the question of how it chooses a specific direction for its migration. The young of many species of birds migrate independently of the adults, setting out on a course they have never traveled before and without the benefit of experienced companions. What causes young, inexperienced birds to select a southerly direction for their first migration? To many investigators the fact that they do so implied that early experience was not important in the development of normal orientation abilities. Some workers even proposed that birds possess a genetically inherited "star map."

Field studies have produced evidence that makes the inherited-star-map hypothesis unlikely. These investigations reveal that there are differences between the navigational abilities of young birds and those of adult birds. When birds of some species, for example the European chaffinch, are captured and displaced from their normal fall migration routes, the adults may correct for the displacement and fly to their regular winter grounds, but young birds on their first migration do not. Prior experience obviously improves the navigational ability of birds.

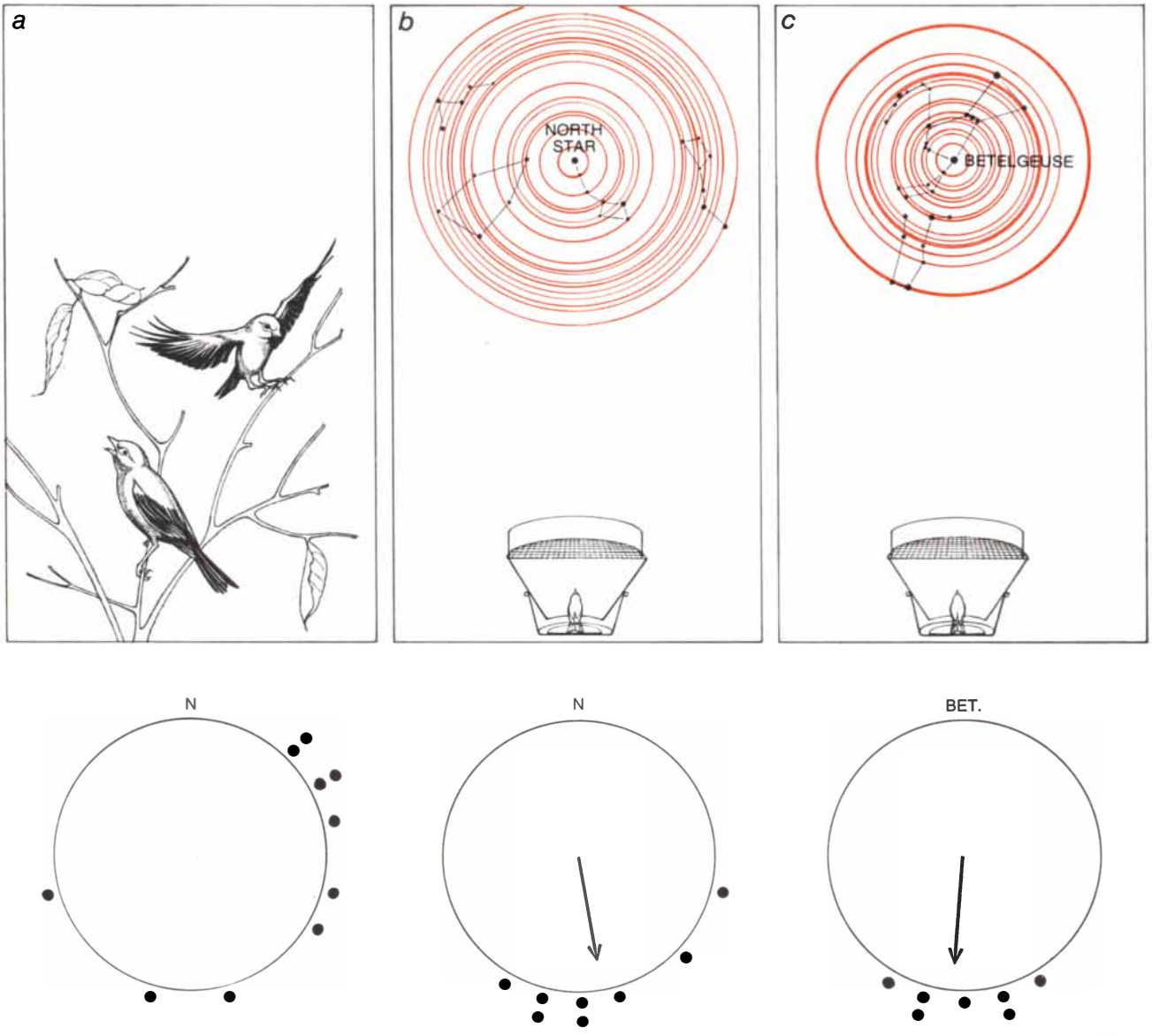
In my studies adult indigo buntings have always been more accurate and consistent in their orientation than young birds. If very young buntings are prevented from the time of their capture from viewing the normal night sky, aberrant orientation behavior develops. One summer I located numerous nests of indigo buntings near Ithaca, and I carefully removed the young birds when they were between four and 10 days old. The nestlings were hand-reared in the laboratory, where their visual experience with celestial cues could be controlled. One group of young buntings never saw a point source of light. They lived in a windowless room with diffuse fluorescent lighting until they began depositing migratory fat and exhibiting intense nocturnal activity in September and October. When I tested these birds in the planetarium under a night sky that matched the normal one for the season, they were unable to select a migratory direction. The young birds were highly motivated, jumping with great frequency in the test cages, but the orientation of the jumps was random. They were unable to obtain directional information from the stars.

A second group of young birds was prevented from seeing the sun, but the birds were allowed to view the night sky in a planetarium every other night during August and September. The star projector rotated at a speed of one revolution every 24 hours, thus duplicating the normal pattern of celestial rotation. When these birds began displaying nocturnal activity, they were tested in the planetarium under the same sky as the group of birds that had had no visual experience with the stars. Unlike the inexperienced birds, the buntings that had already been exposed to the night sky were able to orient to the south. In some way the exposure to the stars was of extreme importance for the normal maturation of star-orientation abilities. Finally,

a third group of young buntings was exposed to the night sky in the planetarium but with one important difference. I modified the star projector by constructing a special arm that allowed the projector to be rotated around any axis of my choosing. I selected Betelgeuse, a bright star in Orion, as the new polestar around which all other stars rotated. The star patterns and constellations remained unchanged; only their positions and movements with respect to the new axis of rotation were altered.

The young buntings of the third group were exposed to the sky with the Betelgeuse polestar every other night for two months. When they became nocturnally active, they were tested under a normal sky with the North Star at the pole position. The buntings consistently oriented their activity 180 degrees away from Betelgeuse—the appropriate "southerly" direction as defined with respect to the polestar of their early experience [*see illustration below*].

On the basis of experiments of this kind I have hypothesized that young buntings respond to the apparent rotational motion of stars in the night sky. The stars near the North Star move through much smaller arcs than the stars near the celestial equator, and this enables the young birds to determine a



EARLY VISUAL EXPERIENCE of young indigo buntings was found to play an important role in the development of their celestial-orientation abilities. Three groups of nestlings were captured and hand-reared in the laboratory. The first group (*a*) lived in a windowless room with diffused lighting and never saw a point source of light. In the fall the birds began to display intense nocturnal activity. When they were tested under a stationary night sky in a planetarium, they were unable to select a migratory direction. The second group (*b*) never saw the sun and was exposed to a night sky in a planetarium every other night for two months. Normal celestial rotation was simulated. When the birds were tested in a planetarium under a normal sky during their fall migratory period, they oriented to the south. The third group (*c*) also never saw the sun and was exposed to a modified night sky in a planetarium every other night for two months. Betelgeuse, a star in Orion, became the new polestar around which all other stars rotated. When the birds were tested in the fall under a normal night sky, they continued to regard Betelgeuse as the polestar and oriented their activity away from it. The experiment shows that young buntings initially learn the north-south axis from the rotation of stars and that star patterns by themselves are not useful cues to a naïve bunting. Star patterns take on directional meaning only after they have become part of the bird's general orientational framework, the formation of which is influenced, at least in part, by observing the rotation of stars.

north-south directional axis. Individual stars and patterns of stars are of no value for direction finding until their positions with respect to some reference framework have been learned. The axis of the rotation of the stars appears to function as one reference system. Once the stellar information and the rotational have been coupled, however, the bunting can locate the rotational axis from star patterns alone. This is suggested by the finding that adult indigo buntings orient accurately even under stationary planetarium skies. Celestial motion thus becomes a secondary or redundant cue for adult birds.

One cannot help but speculate about the possible selective advantage of a maturation process that makes use of celestial rotation for a directional reference system. One possible explanation lies in the long-term unreliability of the stellar cues themselves. The rotation of the earth can be viewed as being analogous to the spinning of a top. And like most spinning tops, the spinning earth wobbles. This slight wobble, usually described as the precession of the equinoxes, causes the direction of the earth's spin axis to shift. Over a period of 26,000 years the precession of the equinoxes causes the earth's spin axis to trace on the celestial sphere a full circle with a radius of 23.5 degrees. This motion gives rise to marked seasonal and latitudinal changes in the apparent position of stars. The spring stars of the present become the fall stars in 13,000 years, and vice versa. The values of declination also change: as the polar axis moves through its circle Vega becomes the new polestar, and the present North Star shifts to 43 degrees north [*see illustration on this page*]. Similar changes occur for all stars.

The possible implications of these changes for the star-navigation system of birds are obvious. If birds were to rely on a genetically fixed star map, the rate of genetic change would have to be extremely rapid to allow for the change in position of the stars. A maturation process in birds that involves finding the north-south axis by the rotation of stars, however, minimizes the problem. Of course, several reference cues may play a role, but the axis of celestial rotation is well suited to function as one such reference because that axis is aligned with geographic north-south regardless of which particular stars and patterns of stars are located near the celestial pole.

Experiments in the planetarium have enabled us to learn a great deal about the orientation of night-migrating birds. Young birds develop a north-south reference axis as a result of early exposure to celestial rotation. Then they learn the patterns of stars around the northern celestial pole, which they use in a configurational manner to determine a direction of migration. The precise direction that is selected depends on the hormonal and physiological state of the bird and not on seasonal differences in the position of the stars. The star-orientation process of the indigo bunting is basically one of pattern recognition that does not involve an internal time sense.

The experiments described here are equally important for what they do not explain about migratory orientation. Although the star-orientation process enables a bird to maintain a given course during its migratory trip, as an explanation for the orientation abilities of birds the process by itself is not entirely adequate and not absolutely essential. I say not entirely adequate because the direction-finding system described for the indigo bunting is basically only a star-compass system that enables the bird to select and maintain a given direction. The system does not provide any information about actual geographic location.

The star compass does not tell the bird it has been blown off course to the east or to the west, nor does it tell the bird when it has reached the latitude of its destination. That is because the star-orientation capability I have described still lacks the map component of the map-and-compass hypothesis. It is quite possible that most migrating birds are not at all goal-oriented during the major portion of their migratory flight. Their process of orientation may be fundamentally different from that of homing pigeons and they may only revert to a homing type of process during the very final stages of the migratory flight.

I say not absolutely essential because we now know that migratory birds have numerous directional cues available to them. In addition to the use of the sun and the stars, experiments have shown that songbird migrants can make use of the position of the sunset, the directionality of the winds aloft, the direction of the earth's magnetic field, the presence of topographic landmarks and the activity and the call notes of other birds of the same species as sources of information, enabling them either to select or at least to maintain a given migratory direction.

Birds thus have access to many sources of directional information, and natural selection has favored the development of abilities to make use of them all. Some cues may give more accurate information than others; some may be available throughout the flight, whereas others may be useful only at specific geographic locations; some may be available regardless of flight conditions, whereas others may be functional only under optimal meteorological situations.

PRECESSION OF EARTH'S AXIS produces change in the apparent position of the stars. In 13,000 years, as the polar axis moves through half of its circle, the star Vega will become the new polestar and the present North Star will shift to 43 degrees north.

The realization that birds have multiple cues at their disposal is in itself a finding of major importance. Although I have concentrated here on the star compass, I do not mean to imply that the other cue systems are not important. The discovery of a hierarchy of redundant directional cues makes the search for a single mechanism of migratory orientation obsolete.

The nocturnally migrating bird should be viewed as an animal whose behavior has been shaped by aeons of intensive selection pressure. It is the combination of the bird's skill as a meteorologist and as a navigator that accounts for its successful traversing of thousands of miles of environmentally inappropriate terrain each fall and spring. Although our understanding of the migratory navigation of birds has come a long way since Kramer began the experimental approach some 20 years ago, the total of our knowledge is still not enough to fully explain how an individual bird finds its way between its breeding territory and its wintering grounds.

MATHEMATICAL GAMES

More about tiling the plane: the possibilities of polyominoes, polyiamonds and polyhexes

by Martin Gardner

"I often wondered at my own mania for making periodic drawings. . . . What can be the reason of my being alone in this field? Why does none of my fellow-artists seem to be fascinated as I am by these interlocking shapes?"

—M. C. Escher

Last month's "Mathematical Games" was devoted to tessellating the infinite plane with congruent, nonoverlapping convex polygons. This month we extend the topic into the vast domain of nonconvex polygons, with special attention to polyominoes, polyiamonds and polyhexes.

Polyominoes, as all regular readers of this department know, are shapes formed by joining unit squares at their edges. They were first introduced to the mathematical world in 1953 by Solomon W. Golomb. His book *Polyominoes* (Scribner's, 1965) is the standard reference on this popular recreation. Some of my previous reports on polyominoes are reprinted in *The Scientific American Book of Mathematical Puzzles & Diversions* and in *New Mathematical Diversions from Scientific American*. Another report on polyominoes, not yet in a book, was published in this department in October, 1965, and polyominoes have played important roles in many other recreations discussed here, such as John Horton Conway's game of Life (October, 1970, and February, 1971).

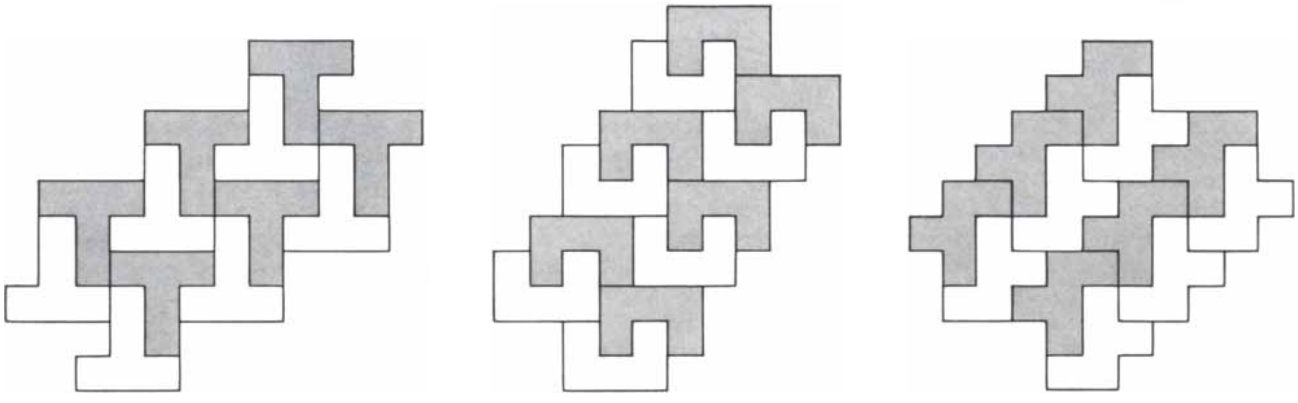
The monomino (single square) and domino obviously tile the plane, and so do the two kinds of tromino. It takes only a few moments to discover that each of the five tetrominoes will tile in a simple periodic pattern in which all the tiles are identically oriented; that is, no tile needs to be rotated or reflected (flipped over).

Each of the 12 five-celled figures will tile the plane. (The term "Pentominoes," by the way, is a trademark registered by Golomb for the games and puzzles involving these 12 geometric shapes.) All but three—the *T*, *U* and *R* pentominoes—tile by simple translation (sliding without rotations or reflections). Each of the three that require rotation can be fitted into pairs by turning one upside down to form an order-10 polyomino that tiles by translation.

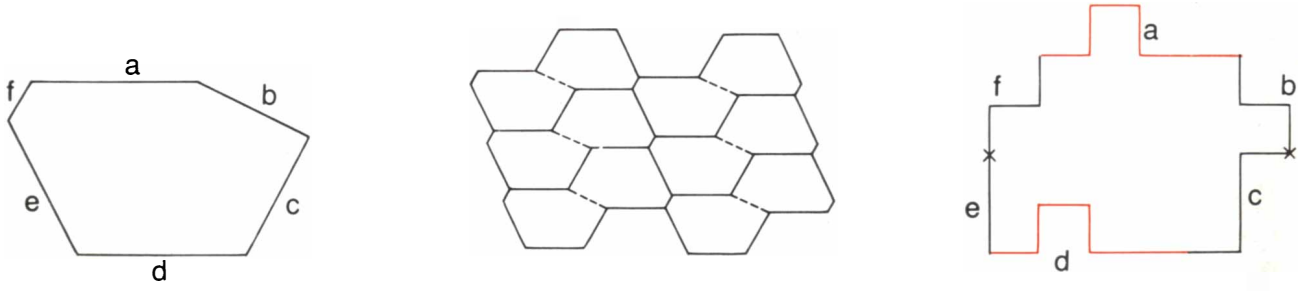
There are 35 distinct hexominoes, and all of them tile without being reflected. Some tile by simple translation. Those that do not can be joined in pairs by turning one upside down to make an order-12 polyomino that tiles by translation.

Some 10 years ago Conway became interested in the tiling properties of polyominoes and other polygons. When he examined the 108 heptominoes, it was clear that the task of identifying the tilers would be tedious unless he could find a criterion that would, as he put it, "dispose of most of them rapidly without diagrams all over the graph paper."

The criterion discovered by Conway is an efficient one that applies to any polygon. It is based on a hexagonal tiling



The three pentominoes that require rotation to tile periodically



Criterion for periodic tiling without reflection

pattern [*see diagrams at left and middle in lower illustration on opposite page*]. Note that edges *a* and *d* are equal and parallel and that throughout the pattern *a* joins *d* with the two hexagons in the same orientation. Note also that each of the other four edges joins its corresponding edge on a tile that has been rotated 180 degrees.

With these facts in mind it is easy to understand the basis of Conway's criterion. We examine a given polygon to see if its perimeter can be divided into six parts, *a, b, c, d, e* and *f*, that meet the following requirements:

1. Two opposite edges, *a* and *d*, are "parallel" in the sense that they are congruent and in the same orientation.

2. Each of the other four edges, *b, c, e* and *f*, are centrosymmetric; that is, they are unaltered by a 180-degree rotation around a midpoint.

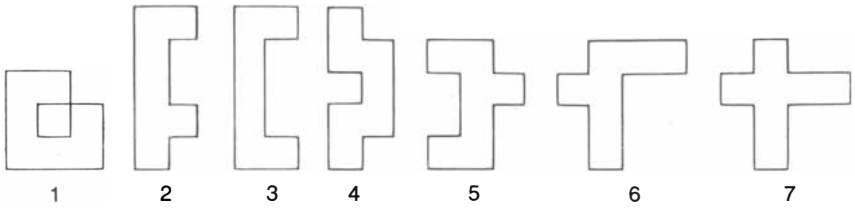
If the polygon meets both requirements, it will tile the plane periodically. No flipping over of tiles is necessary. The tiles are paired simply by turning one upside down to form a figure that tiles by translation.

Conway's example will make this procedure clear [*see diagram at right in lower illustration on opposite page*]. The two colored lines mark sides *a* and *d*, which are "parallel." The two *X*'s distinguish the edges *b, c, e* and *f*. Each of the four edges has central symmetry; therefore the figure will tile the plane by pairing. One figure of the pair is rotated 180 degrees with respect to the other, without reflection, and the double shape tiles by translation. It is important to realize, Conway adds, that any of the six edges may be empty (nonexistent). The criterion is quite general, applying to triangles, quadrilaterals, pentagons and all higher polygons.

Of the 108 heptominoes, 101 meet Conway's criterion. That means each can be paired with an upside-down mate to form a 14-cell shape that tiles by translation. The seven that fail to meet the criterion are shown in the top illustration on this page. The first heptomino obviously cannot tile the plane because there is no way to fill the hole. The fourth one tiles by pairing with a 90-degree rotation as is shown in the illustration at left at the bottom of this page.

The fifth, which Conway considers the most interesting heptomino, tiles in two ways. It tiles by pairing, with 90-degree rotation and reflection. Without reflection it tiles in quadruplets of four orientations [*see middle illustration*].

The second heptomino in this group will not tile without reflection. The smallest region, which tiles by translation, contains four replicas in four orientations; two of the replicas are reflected, as is shown in the illustration at the lower right below. This is the lowest-order polyomino, unique among the heptominoes, that requires reflection to tile the plane.



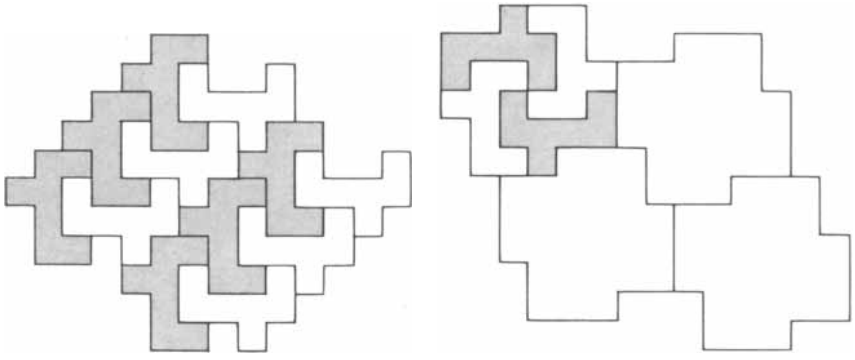
The seven heptominoes that do not meet the criterion for tiling periodically

The third, sixth and seventh heptominoes will not tile, making four nontilers in all. Proving impossibility in each case is not difficult, although it is sometimes tedious. First you try all possible ways of fitting two replicas together, eliminating the ways that have a hole or a space that cannot be filled by a third replica. After all such pairs have been found you test all ways of adding a third piece that allow the placing of a fourth. Eventually you reach a set of *n* heptominoes such that, no matter how they are combined, there is no way to add another replica. I leave these proofs to interested readers.

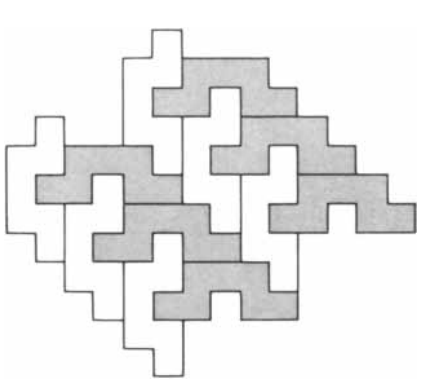
Although the third heptomino will not tile the plane, it can be combined with 3-by-3 square tiles to form a striking tessellation. The reader is urged to make a set of cardboard replicas of this heptomino, together with replicas of the order-3 square, and see if he can combine the shapes to form a tessellation. The pattern will be given next month.

As far as I know only David Bird of North Shields in England has isolated the nontilers among the 369 octominoes. The six with holes can of course be eliminated immediately. Bird reports 20 other nontilers, making 26 in all [*see upper illustration on next page*]. I should welcome either confirmation or counterexamples.

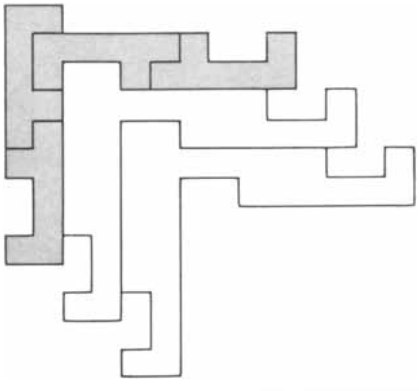
Is there a general algorithm that can be applied to any polyomino to decide if it will tile? Conway's criterion identifies certain polyominoes as tilers, but if a



Periodic tiling of heptomino 5



Heptomino 4 tiles with 90-degree rotation



Heptomino 2 tiles only with reflection

polyomino does not meet his criterion, it may or may not tile in other ways. Conway conjectures that there is no general algorithm, but the conjecture has not yet been proved. A major step toward establishing the undecidability of the algorithm was reported by Golomb in his paper "Tiling with Sets of Polyominoes," in *Journal of Combinatorial Theory,* Vol. 9, pages 60–71, July, 1970. In that paper Golomb considers whether there is a procedure for deciding if any given finite set of different polyominoes (assuming an unlimited supply of each kind) will tile the plane. He shows that the problem is equivalent to a problem of tiling the plane with a finite set of edge-colored squares by matching colors at the joined edges. Since Hao Wang and his colleagues had earlier proved the latter problem to be undecidable (see "Games, Logic and Computers," by Hao Wang; SCIENTIFIC AMERICAN, November, 1965) the corresponding problem for polyominoes is also undecidable.

Golomb's paper extended the results published in an earlier article, "Tiling with Polyominoes," in *Journal of Combinatorial Theory,* Vol. 1, pages 280–296, September, 1966. In his first paper Golomb considers the tiling of such subsets of the plane as the half-plane, the quarter-plane, straight strips, bent strips and rectangles. A chart summarizes results through the hexominoes.

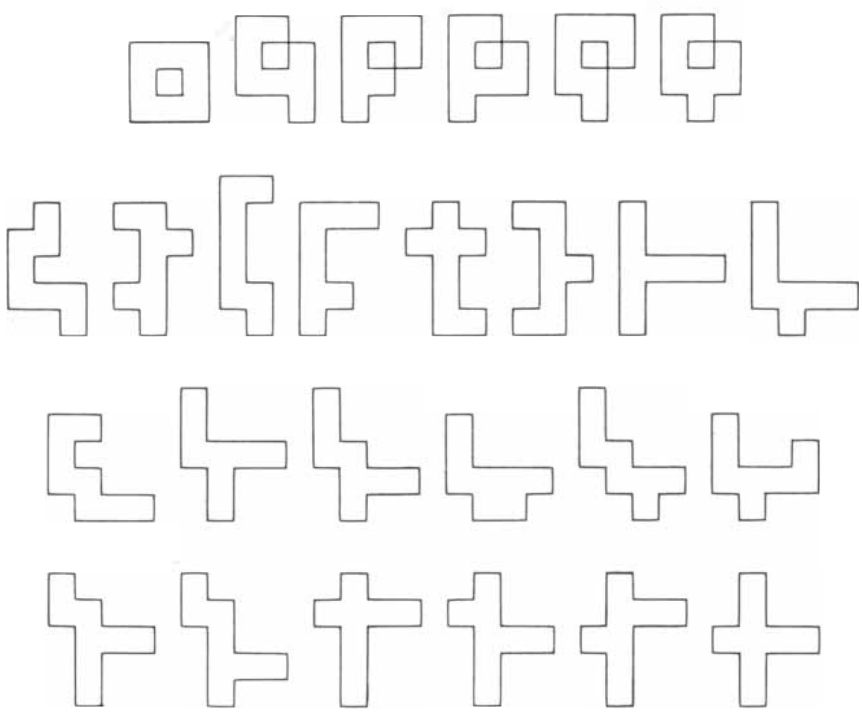
If there is no region in a tessellation that tiles by translation, the tessellation is said to be nonperiodic. One curious way of tiling nonperiodically with congruent polygons is to use a set of them to build a larger replica. Obviously sets of these larger replicas can then be combined in the same way to make a still larger replica, and in that way the entire plane can be covered. A tile with the property of self-replication was named by Golomb a "rep-tile." (See his paper "Replicating Figures in the Plane" in *The Mathematical Gazette,* Vol. 48, pages 403–412, December, 1964, and the chapter in *New Mathematical Diversions* cited earlier.)

Less work has been done on polyiamonds (figures formed by joining congruent equilateral triangles) than on polyominoes. (On the polyiamonds see Golomb's book and chapters 18 and 24 in *Sixth Book of Mathematical Games from Scientific American,* now available in paperback from Scribner's.) It is not hard to establish by Conway's criterion that all polyiamonds through order 6 will tile. Of the 24 heptiamonds only the *V* heptiamond is not a tiler. (A simple impossibility proof is given in Chapter 24 of *Sixth Book*). Bird, Gregory J. Bishop, Andrew L. Clarke, John W. Harris, Wade Philpott and others have found that all octiamonds will tile. As far as I know the 160 enneiamonds (order 9) have not yet been settled, although two correspondents, Bird and Clarke, agree that the nontilers are the 21 shown in the lower illustration at the left, based on diagrams supplied by Bird.

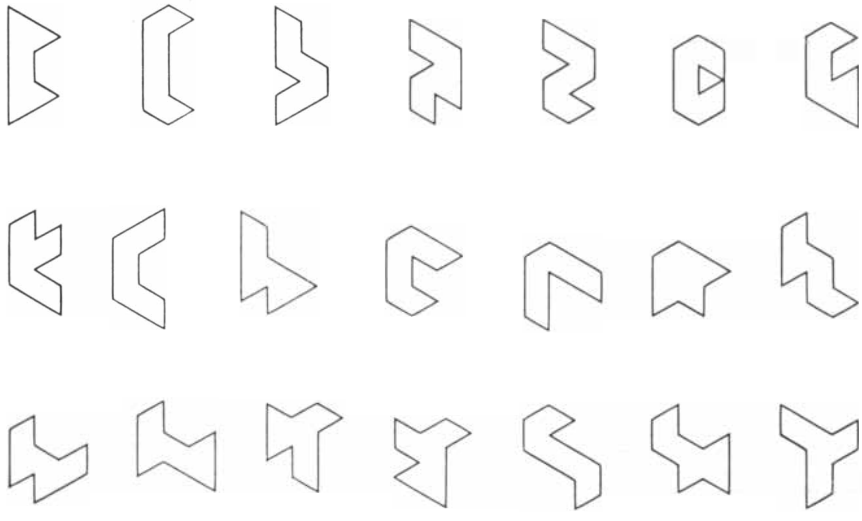
The polyhexes, formed by joining congruent regular hexagons (see "Mathematical Games" for June, 1967), have received even less attention than the polyiamonds. Bird and Bishop have established that all polyhexes through order 5 are plane-fillers. Of the 83 hexahexes Bird believes only five are nontilers [*see top illustration on opposite page*].

Once it is determined that a given polygon will tile one can ask in how many distinct ways the tiling can be accomplished. It can be a very sticky question. A. W. Bell, in his paper "Tessellations of Polyominoes," in *Mathematical Reflections,* edited by members of the Association of Teachers of Mathematics (Cambridge University Press, 1970), has made a beginning by attempting to classify all patterns for the *L* tetromino. He obtained 19 patterns but made no claim for completeness.

Major P. A. MacMahon's little book



The 26 nontiling octominoes



The 21 enneiamonds believed to be nontilers

The five hexahexes thought to be nontilers

New Mathematical Pastimes (Cambridge University Press, 1921; why has no publisher reprinted it?) explains with a wealth of illustrations how a simple polygon tessellation is easily transformed into a more complicated one. You merely take a pair of straight edges, which always go together in the pattern, and change the straight boundary to a crooked one (curves are allowed), subject to certain symmetry restrictions. It was by just such techniques that Islamic craftsmen created the intricate abstract mosaics in the Alhambra and the Taj Mahal.

Here is how Escher described the difficulty of adapting the technique to the creation of tiles that resemble living things: "The borderline between two adjacent shapes having a double function, the act of tracing such a line is a complicated business. On either side of it, simultaneously, a recognizability takes shape. But the human eye and mind cannot be busy with two things at the same moment, and so there must be a quick and continual jumping from one side to the other."

Tessellation theory, quite apart from its usefulness to artists who design patterns for walls, floors, fabrics and so on, has a practical application in industry. In cutting congruent shapes from thin sheets of metal, plastic, cardboard, leather and other materials, a tessellation pattern obviously provides the only way of doing it without waste. An unusual art book could be made by collecting pictures of the tessellations used in modern manufacturing, from the simple rectangular patterns of postage stamps, dollar bills and playing cards to complicated machine parts.

There is also a potential application to jigsaw puzzles. In 1958 the British mathematician Roger Penrose amused himself by transforming a parallelogram tessellation into tilings of order-18 polyiamonds. Some of them make excellent puzzles. Consider a "loaded wheelbarrow" polyiamond [*see illustration at right*]. It is given here for the first time with Penrose's permission. If the reader will make 12 cardboard replicas of the figure, he will find it a challenging task to fit them together to make a region that tiles the plane by translation. The pattern will be disclosed next month.

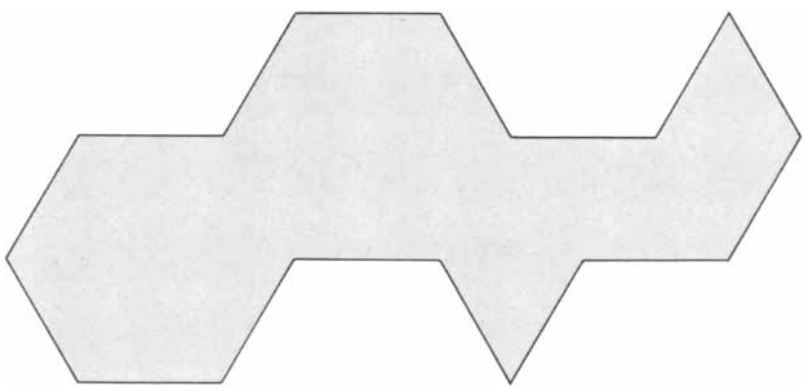
Penrose is Rouse Ball Professor of Mathematics at the Mathematical Institute of the University of Oxford. He is best known among physicists for his contributions to relativity theory and cosmology. He and his father, the late L. S. Penrose, were the first to discover "impossible objects" such as the famous Penrose staircase that Escher used so effectively in his lithograph *Ascending and Descending*.

For about a decade it has been known that there are sets of polygons that together will not tile the plane periodically but will do so nonperiodically. A few years ago Raphael M. Robinson constructed a set of six tiles that tile only nonperiodically. Penrose later found a set of four and finally a set of just two. Is there a single shape that tiles only nonperiodically? We know it cannot be convex (from R. B. Kershner's results given last month), but whether there is a nonconvex tile with that property is one of the deepest unsolved problems in tessellation theory. A rep-tile that cannot also tile periodically would provide such a shape, but so far no such animal has been discovered. The subject of nonperiodic tiling is, however, another story, and one I hope to discuss in some future article.

When I wrote briefly about Hadamard matrixes in the May issue, I had not realized that their application to telemetry codes was discovered by Golomb and that the use of such codes in several Mariner probes of Mars was the result of Golomb's affiliation with the Jet Propulsion Laboratory of the California Institute of Technology. (See "The Search for Hadamard Matrices," by Golomb and L. D. Baumert in *The American Mathematical Monthly*, Vol. 70, pages 12–17, January, 1963.) A Hadamard matrix larger than 2 by 2 must have a side that is a multiple of 4, but it is not yet known if matrixes exist for all such orders. I reported order 188 as the first unsolved case. Golomb tells me that an order 188 was recently found by Richard J. Turyn and that the lowest unknown case now is 268. Baumert, at the Jet Propulsion Laboratory, is keeping track of all cases under 1,000.

Hadamard matrixes have also found extensive application in the processing of pictorial information. The "Hadamard transform" (analogous to the "fast Fourier transform") actually produces a mathematical hologram of the original image. Golomb has been a pioneer in recognizing that classic combinatorial design problems often provide optimum solutions to data-processing problems and that it pays to look for engineering problems to which these designs are the solution.

The second solution reported in June to the chess problem that I first described in March has proved to be erroneous. Many readers have written to say correctly that the mate is easily thwarted by Black's P-KN4. The original solution is therefore unique.



Roger Penrose's polyiamond puzzle

THE AMATEUR SCIENTIST

Graphs that predict when planets will line up with another planet or the sun

Conducted by C. L. Stong

From time to time a planet, as it is observed from the earth, moves into alignment with the sun or with another planet. These alignments are termed conjunctions or oppositions, according to the relative positions of the earth and the sun. They can be predicted by a graphical technique that has been devised by B. E. Johnson (2233 81 Avenue SE, Mercer Island, Wash. 98040).

Conjunctions and oppositions have attracted interest throughout recorded history. To astrologers they were portents, usually dire. Today the forces that arise from such groupings are taken into account by digital computers to help determine the course of space probes.

To amateur observers conjunctions and oppositions form interesting patterns in the night sky. Moreover, a conjunction marks the onset of spectacular motions that appear to carry the planets involved along a looping path against the background of the fixed stars. The looping motion can be followed by making nightly observations for a few weeks.

The mathematical prediction of planetary alignments can be tedious, even with the aid of a pocket calculator. No graphical technique can yield an exact prediction. Moreover, planetary motions are not exactly proportional to the passage of time. Even so, Johnson's method generates results of sufficient accuracy to satisfy most amateur needs. Johnson, who is an electrical engineer, describes his scheme as follows.

"Planetary alignments are of four kinds, depending on the distance of the planets from the sun. The two 'inferior' planets, Mercury and Venus, are closer to the sun than the earth is. Mercury completes one orbit around the sun in 87.969 earth-days; Venus, in 224.7 earth-days. These intervals are the sidereal periods of the planets. (The sidereal period of the earth is 365.2563 days. The interval between two successive vernal equinoxes of the earth is the tropical year. It spans 365.2422 days.)

"The sidereal period of each planet increases with the planet's distance from the sun. The inclinations of the orbital planes of the planets differ from one another by only a few degrees, with the exception of the orbital plane of Pluto. The orbit of Pluto is inclined about 17 degrees to the ecliptic, the plane in which the earth travels.

"An apparent alignment of the inferior planets occurs when Mercury or Venus moves exactly between the earth and the sun. (Sometimes both of them move.) These events are called inferior conjunctions. Continued rotation carries each planet at its characteristic velocity to the opposite side of its orbit to points where the sun lies in the line between the planets and the earth. These positions are termed superior conjunctions.

"The 'superior' planets (Mars, Jupiter, Saturn, Uranus, Neptune and Pluto) are, in that order, increasingly distant from the sun and have correspondingly longer sidereal periods. A superior planet reaches conjunction when it moves to the point where the sun lies between the planet and the observer's meridian at noon. At that position the planet is lost from view in the sun's glare.

"Continued rotation carries each superior planet to the point where the earth is eventually aligned between the planet and the sun. The planet is then said to be in opposition. The position appears on the observer's meridian at midnight. Periodically several superior planets reach opposition simultaneously. Inferior planets can never reach opposition, and superior planets cannot reach inferior conjunction [*see top illustration on opposite page*].

"The interval of time between successive conjunctions or successive oppositions of a planet is its synodic period. This period can be calculated easily by either of two methods. For a superior planet the reciprocal of the synodic period is equal to the difference between the reciprocal of the earth's sidereal period and the reciprocal of the planet's sidereal period. With Mars as an example the reciprocal of the synodic period is $1 / 365.25 - 1 / 686.98 = 1 / 779.9$. The interval between successive conjunctions or successive oppositions of Mars is 779.9 earth-days.

"The synodic period of the inferior planets is given by the difference between the reciprocal of the planet's sidereal period and the period of the earth. For Venus the reciprocal of the synodic period is $1 / 224.7 - 1 / 365.25 = 1 / 583.94$. The interval between successive superior or inferior conjunctions of Venus is 583.94 earth-days.

"A simple graph can also display the synodic period of a planet [*see bottom illustration on opposite page*]. Draw a pair of rectangular coordinates. Calibrate the abscissa in intervals of one year for, say, 2½ years and the ordinate in units of 360 degrees for at least 720 degrees. The fact that the earth completes an orbit around the sun (traverses 360 degrees) in one year can be represented by a graph in the form of a diagonal line drawn from zero on the abscissa to intercept 360 degrees at one year. The slope of the graph is equal to the quotient of 360 degrees divided by the number of days in a year, or .9856 degree per day.

"The graph of any planet can be similarly plotted on the same coordinates. For example, the motion of Mars can be represented by a straight line that begins at zero on the abscissa and intercepts the line representing 360 degrees at the sidereal period of Mars (686.99 earth-days). The slope of the Martian graph is equal to the quotient of 360 degrees divided by the number of earth-days in one sidereal year of Mars, or .5240 degree per day. The difference in the rate at which the two planets travel around the sun is indicated by the difference in the slopes of the graphs.

"Periodically the difference between the graphs as measured on the ordinate amounts to a multiple of 360 degrees. The multiples occur at time intervals equal to the synodic period of the planet and are equal to 360 degrees divided by the difference between the slopes of the

graphs: 360 / (.9856 − .5240) = 779.9, which is the synodic period of Mars. Graphically the synodic period can be measured along the abscissa. It amounts to 2.135 sidereal earth-years.

"A simple modification can adapt this graphical technique for displaying both past and future conjunctions and oppositions. Limit the ordinate to one interval of 360 degrees. Divide the abscissa into any number of yearly intervals. The accuracy of the predictions increases with the size of the graph.

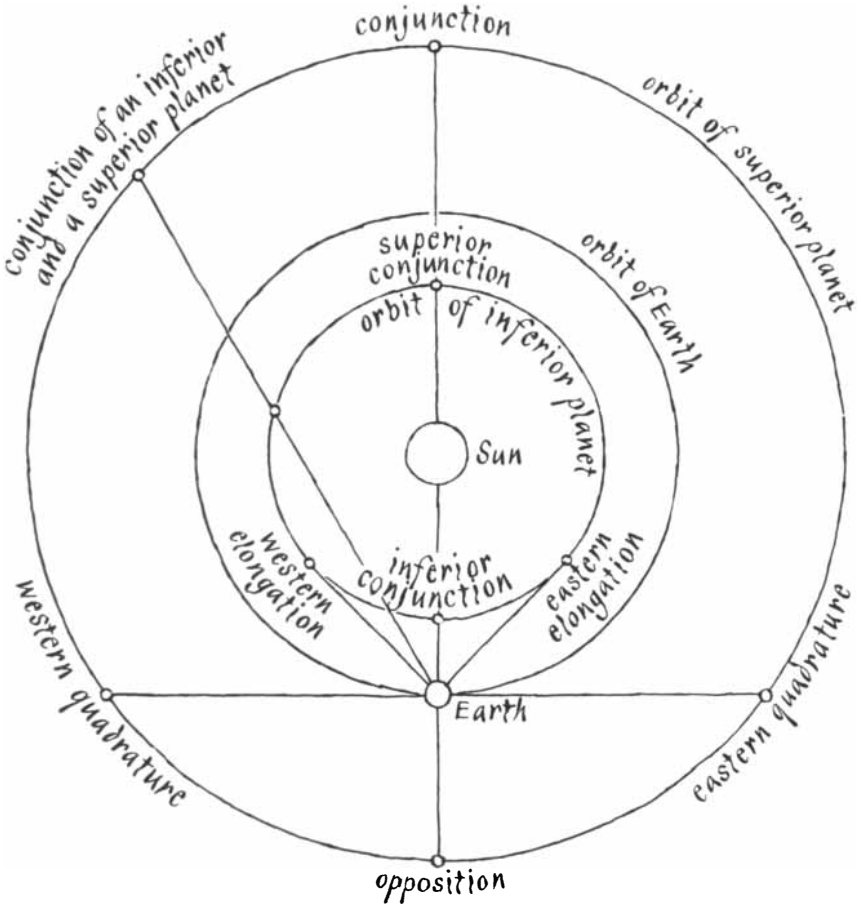
"The graph of each planet originates on the abscissa at the beginning of the sidereal period and terminates at the point where the end of the period intercepts 360 degrees as measured by the ordinate. Continued motion of the planet is depicted by initiating a new graph at zero degrees and at the first day of the next sidereal interval as suggested by the broken line 'Earth'' and the solid line 'Mars'' in the illustration. The result is a sawtooth pattern on the graph of each planet.

"The slope of the 'teeth' and their number vary with the sidereal period of each planet. The accompanying graph [*upper illustration on next page*] depicts motions of the earth and Mars through the decade ending in 1979. The ordinate of the graph is divided into 12 intervals of 30 degrees. Each interval represents the partial orbit of the planet around the sun in roughly one month. Each interval of 30 degrees is labeled to indicate approximately the constellation in which the sun appears at the time and also the corresponding months during which the constellations appear on the meridian.

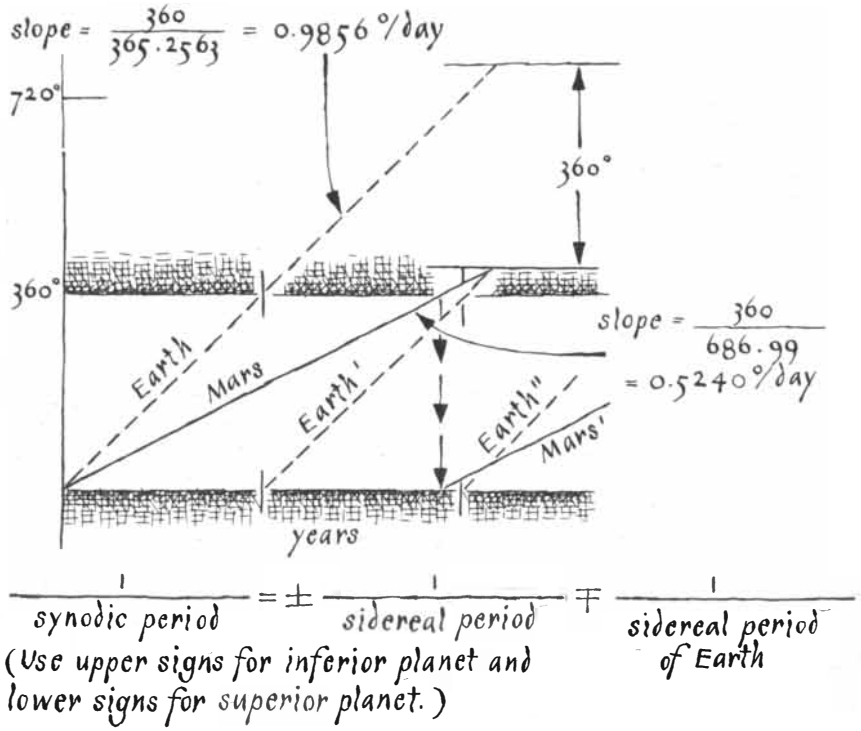
"The abscissa of the graph is divided into 10 intervals that represent the years 1969 through 1978. To avoid cluttering the graph with fine detail the yearly intervals have not been subdivided into intervals representing months. A black line that represents the motion of the earth for 1969 is drawn on the graph beginning on January 1, 1969, at zero degrees and sloping diagonally to December 31 at 360 degrees.

"A colored line begins at approximately 75 degrees on the ordinate and January 1 on the time scale. This graph depicts the motion of Mars. It intercepts the time scale approximately 548 days later (about July 2, 1970). On that date the new graph of Mars begins at zero degrees, as measured along the ordinate. In contrast, graphs of the earth's motion begin and end with each calendar year.

"Note the periodic intersections of the two lines every 779.9 earth-days. The intersections mark the dates when Mars is in opposition, as it will be on Decem-



Orbits, conjunctions and oppositions

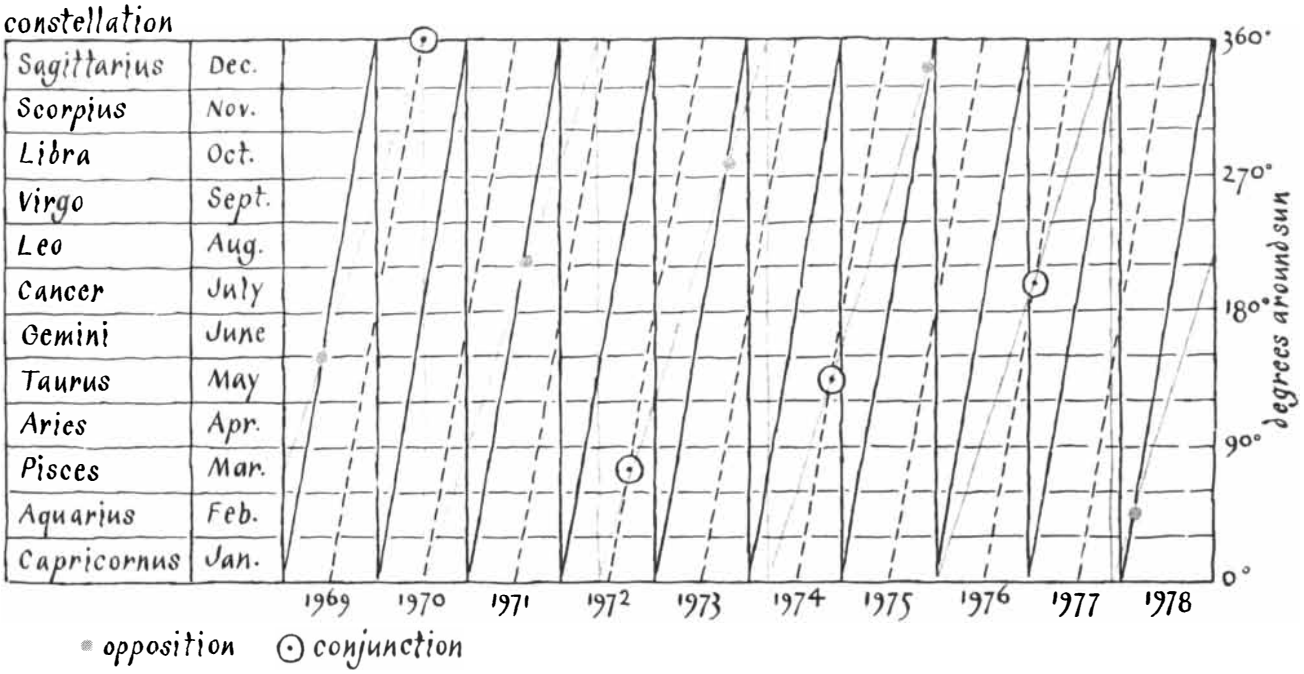


Basic graph of planets in orbit

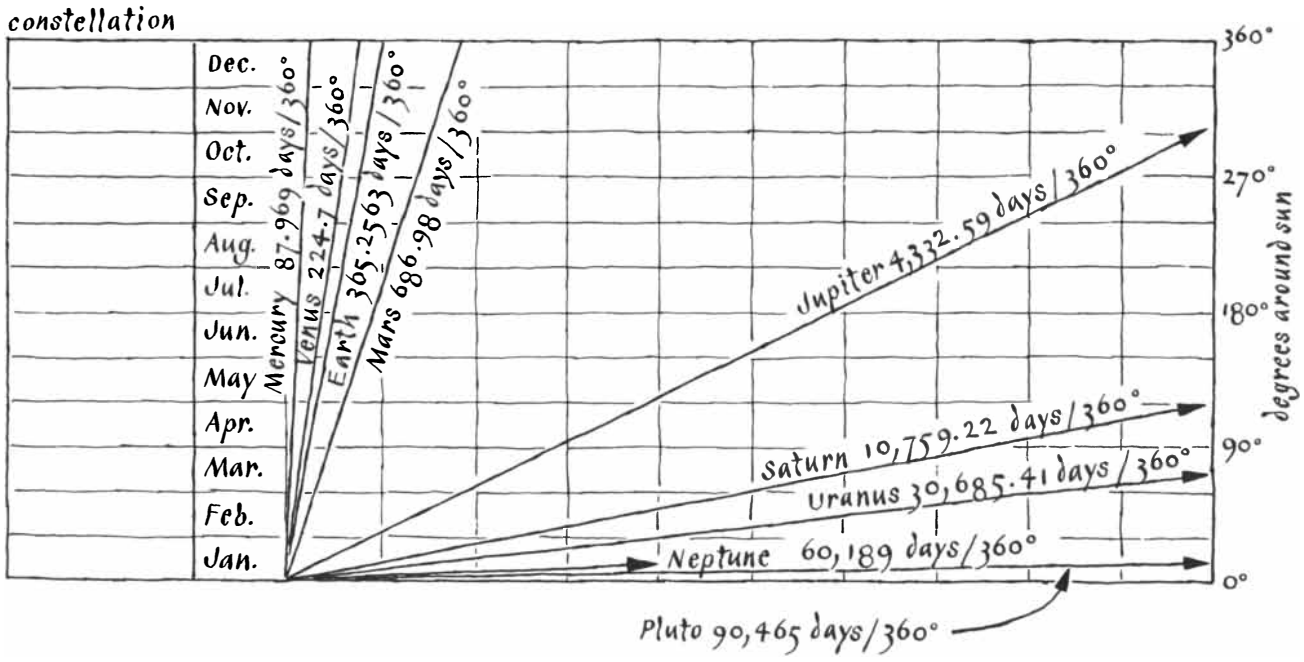
ber 15, 1975. Conjunctions can be depicted by displacing the graph of the earth's motion by an interval of six months with respect to the abscissa, as indicated by the broken line. Dates when Mars has been or will be in conjunction are indicated by the intersections of the broken and the solid lines.

"As a convenience I rule a pattern of horizontal and vertical lines 9¾ by 4½ inches in size on standard 11-by-8½-inch paper and label the 12 horizontal lines with the months in the second column from the left [*see lower illustration below*]. At the right margin the same 12 lines are labeled in multiples of 90 degrees from zero through 360 degrees. Duplicates of this form are made on an office copying machine. The accompanying drawing of the form depicts the graphs of the nine planets. It illustrates how the display tends to become crowded with planets within 200 million miles of the sun as well as those that are on the order of a billion or more miles away.

"The slope of any graph of this kind can be plotted most accurately by locating the origin at zero degrees and zero years and terminating it either at the intersection of the planet's sidereal period and 360 degrees or at the point on the orbit (expressed in degrees) that the planet would reach within the maximum number of years displayed by the abscissa. In the accompanying example the slopes of the graphs of Mercury, Venus,



B. E. Johnson's graphical way of displaying conjunctions and oppositions



Relative slopes of all planets

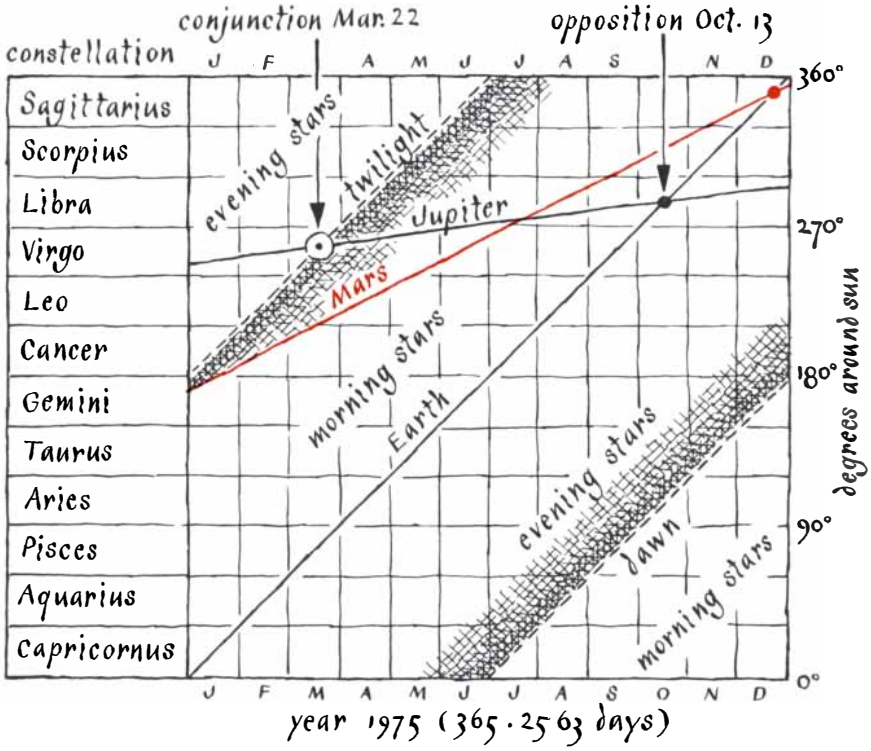
the earth and Mars were determined by drawing straight lines between the point of origin and the 360-degree points that corresponded to the respective sidereal periods of the four planets within 200 million miles of the sun.

"In this example the abscissa spans 10 years, or 3,652.5 days. To what point along the right margin of the chart should a graph be drawn to represent the slope of Pluto? The sidereal period of Pluto is 90,465 earth-days. In one day Pluto would rotate 360 / 90,465 = .0039794 degree. In 10 years it would orbit through 3,652.5 days multiplied by .0039794 degree per day, or 14.53 degrees. A straight line drawn from this point on the right margin of the coordinates to the point of origin at the left displays the slope of Pluto's graph.

"Because the synodic period is a function of the sidereal period, I calibrate the coordinates of the graphs in terms of the sidereal year of 365.2563 days and initiate the orbit at zero degrees on January 1 instead of at the vernal equinox, as would be done in the case of the tropical year of 365.2422 days. With the slopes determined, the graphs of the planets must then be displaced in time from the point of origin as required to locate the intersections of the planet and earth graphs at the exact dates of known conjunctions and oppositions. Almanacs list the dates of these events for each calendar year. A reference that is remarkably complete and easy to use is *The American Ephemeris and Nautical Almanac.* The volume is updated and reissued annually. It can be obtained for $6.20 postpaid from the Superintendent of Documents (U.S. Government Printing Office, Washington, D.C. 20402). The British edition, *The Astronomical Ephemeris,* is published by Her Majesty's Stationery Office and is available from 49 High Holborn, London W.C. 1.

"Having plotted graphs of the earth's motion for as many yearly intervals as are provided by the coordinates, I put one or more points on the graphs to indicate oppositions of the planet to be plotted. The graph of the planet is drawn through these points at the slope previously determined for the planet. A line that has been drawn lightly from the zero point to determine the slope of a planet can be conveniently transferred to intercept the graph of the earth at points of known conjunctions or oppositions by the instrument known to navigators as the parallel ruler.

"Planets can be in conjunction with each other as well as with the sun. These events appear as intersections between the graphs of the planets at points removed from the graph of the earth. For example, the accompanying graph [*above*] indicates that Mars and Jupiter were in conjunction about July 2. Because their intersection appears above the graph of the earth they were seen as morning stars. Evening events appear below the graph of the earth's position."



Prediction of morning and evening events

The history of technology is full of instances in which more than one person makes the same invention at about the same time. A recent example is the Princeton "sail wing," an efficient airfoil of light and flexible construction that has found application in hang gliders and sailboats [see "The Amateur Scientist," SCIENTIFIC AMERICAN, December, 1974]. The effectiveness of the sail wing in powering small boats is indicated by the following letter from a yacht designer, Peter Hodgins (248 Roger Road, Ottawa, Ontario, Canada, K1H 5C6).

" 'Blow me down!' to coin a phrase, if the top drawing on page 141 of your December 1974 issue does not show a sail-wing configuration almost identical with the mainsail I designed and rigged on a dinghy. Based on its performance, I have named the dinghy *Dreamboat.* Nat-

PLANET	ANGULAR VELOCITY IN ORBIT (DEGREES PER EARTH-DAY)	SIDEREAL PERIOD (EARTH-DAYS)	SYNODIC PERIOD (EARTH-DAYS)
MERCURY	4.092339	87.96	115.88
VENUS	1.6021	224.70	583.92
EARTH	.985609	365.25	——
MARS	.524033	686.98	779.94
JUPITER	.083091	4,332.59	398.88
SATURN	.033460	10,759.22	378.09
URANUS	.011732	30,685.40	369.66
NEPTUNE	.005981	60,189.00	367.49
PLUTO	.003979	90,465.00	366.73

Table of planetary motions

SCIENCE/SCOPE

Telesat Canada launched its third Anik satellite in May to handle the rapid expansion of telephone, television, and radio service to Canada's northern communities. Its current network of 50 earth stations will be expanded to 70 by mid-year. Compact, transportable earth stations are contributing to the success of oil exploration crews by providing direct telephone service via the Hughes-built Anik satellites to company headquarters and workers' families.

Two Anik-type satellites and 10 earth stations, now being built by Hughes for the Republic of Indonesia as part of a total telecommunications system, will link the 5,000-island nation with telephone, telegraph, television, and teletype.

From the launching of Early Bird 10 years ago to the six Intelsat IV satellites that now encircle the world, transoceanic telephone calls have increased from three million to more than 50 million. In the same period, the cost of a call from the U.S. to Europe has been cut in half. Hughes built both Early Bird, world's first commercial synchronous communications satellite, and the Intelsat IVs for Comsat Corporation, manager of services for the 89-nation International Telecommunications Satellite Organization.

To handle the 200 million transoceanic calls forecast for 1980, Hughes is now building a new series of satellites -- the Intelsat IV-As -- which will have nearly double the capacity of the present Intelsat IVs.

How technology can offset inflation is illustrated by the 25-year record of missile manufacturing at the Hughes Tucson plant. In the early 1950's, the U.S. Air Force air-to-air Falcon -- most advanced of its day -- cost about $100 per pound. Today, Hughes/Tucson delivers about 16 tons of U.S. Army anti-tank TOW, U.S. Navy air-to-air Phoenix, and U.S. Air Force air-to-ground Maverick missiles each day at an average cost of less than $50 per pound. In fact, Maverick -- which seeks out its targets with a tiny nose-mounted television camera -- costs just pennies more than $25 per pound.

Hughes needs satellite communications engineers: 1) BS/MS EE with experience in design of satellite ground control systems, unit design, testing, and integration; telemetry, command, and ranging system experience desirable. 2) BS/MS mechanical engineer with experience in tracking, telemetry, and command equipment design, on-site testing, and product or packaging design of earth station equipment. U.S. citizenship required. Send resume to: Hughes Aircraft Co., P.O. Box 92919, Los Angeles, CA 90009, Atten. Allan Z. St. Jacques. Equal opportunity M/F employer.

A new device for limiting short circuits in AC power systems is being developed by Hughes to meet the electric utility industry's need for more efficient and compact high-voltage transmission equipment. Hughes is building three current-limiting devices (CLD) for the American Electric Power Service Corporation. The CLD prevents short-circuit currents from reaching unmanageable or destructive magnitudes by rapidly inserting a current-limiting resistance into a short-circuited line.

urally I thought the sail wing was my own invention. It is too bad for me that I was not the first with the principle. On the other hand, it was great to see in your magazine the wind-tunnel test results that I suspected but could never afford to make.

"My sail differs in a number of details from the wing of the hang glider you described in 1974. Essentially the details involve the way in which the basic principle is exploited. For example, because of the competitiveness of the sailboat market I was obliged to simplify the cut of the sails, yet the rigging details had to be arranged in such a way that the sail could be hoisted and lowered easily by a halyard while the boat was under way.

"As yachtsmen know, the gradient of wind speeds and the direction of the apparent wind differ along a sail's height. To compensate for these differences the head of the sail should be allowed to sag off to leeward with respect to the position of the boom at the foot. In other words, it is generally agreed that the sail should have some twist. For this reason the control mechanism used in *Dreamboat*'s mainsail luff, or leading edge, is set slightly loose to allow for the seeming difference in angle of attack. The design maximizes forward thrust. Drag to leeward is minimized.

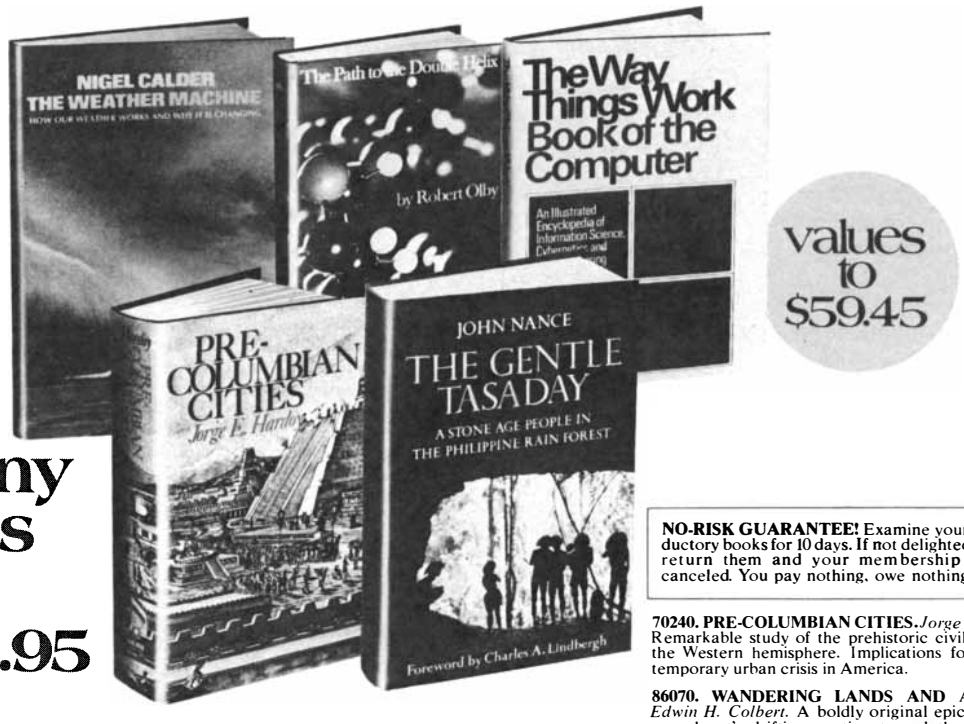
"Another aerodynamic refinement that helps to account for the dinghy's remarkable performance, which is rare in such a craft, is a masthead jib with ample vision below. A light wood fitting at the peak of the mast functions as an end plate that minimizes the vortex normally shed by the sail. The end plate thus minimizes induced drag that would aggravate the tendency of the boat to heel, or list to leeward. Somewhat similar panels, which perform much the same function at the foot of the mast, are sewn into both the mainsail and the jib.

"Three halyards (for mainsail, jib and spinnaker) are placed between the two surfaces of the sail wing to minimize drag. The halyards could have been run inside the aluminum mast. Instead the mast was sealed to provide buoyancy in the event the boat capsized. The wood end plate at the head of the mast contributes still more buoyancy.

"In spite of these attractive features, sales of the boat have not been large compared with sales of traditional craft in the same price range. Yachtsmen tend to be bound by tradition. The craft somehow looks 'different.' Therefore I am designing a new model with an improved hull that will combine the sail wing with a more traditional appearance."

Early man... the weather... new ways with computers... a new look at the double helix...

to keep in touch with everything that's happening in every scientific area that interests you, get set now for a great reading spree:



values to $59.45

Take any 3 books all for only $3.95

as a member of The

Library of Science

(if you will join now, and accept only 3 more books within the next 12 months).

65235. OF TIME, PASSION, AND KNOWLEDGE. *J. T. Fraser.* One of the world's leading authorities in the interdisciplinary study of time offers a wideranging overview of the concept of time. *Counts as 2 of your 3 books.* **$20.00**

50220. THE GENTLE TASADAY. *John Nance.* A veteran journalist's first-hand account of how these primitive food-gathering people were found, what scientists have discovered about their culture, and what the future holds for them. **$15.00**

The Library of Science 2-1AB
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $3.95 for all three. I agree to purchase at least three additional Selections or Alternates during the first 12 months I am a member, under the club plan described in this ad. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments. **Send no money. Members are billed when books arrive.**

3 books for $3.95

Indicate by number the 3 books you want.

[ ] [ ] [ ]

A few expensive books (noted in book descriptions) count as 2 choices.

Name _____

Address _____

City _____ State _____ Zip _____

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.) Book selections purchased for professional purposes may be a tax-deductible expense.

(Publisher's Prices shown.)

39950. COMPUTER AIDED EXPERIMENTATION. *Jules Finkel.* Lays out the basic concepts and techniques needed to specify, design, and use minicomputers in experimentation. *Counts as 2 of your 3 books.* **$24.95**

86540. THE WAY THINGS WORK BOOK OF THE COMPUTER. A veritable encyclopedia of information science, cybernetics, and data processing for the non-specialist. **$8.95**

68690. THE PHILOSOPHY OF QUANTUM MECHANICS. *Max Jammer.* This 500-page volume explores virtually all of the major interpretations— from semi-classical through the hidden variable theories, to the modern stochastic and statistical interpretations. *Counts as 2 of your 3 books.* **$19.95**

34720. ATLAS OF ANCIENT ARCHAEOLOGY. *Edited by Jacquetta Hawkes.* First sourcebook pinpointing the world's major archaeological sites, with 350 maps, site plans and drawings. The combined work of almost two dozen internationally renowned archaeologists. *Counts as 2 of your 3 books.* **$19.50**

62630. THE MILKY WAY. Fourth Edition. *Bart Bok and Priscilla Bok.* Updated and expanded outsized volume, filled with over 125 pictures, charts and diagrams, includes the latest scientific research in wonderfully elegant style. **$15.00**

73330. A RANDOM WALK IN SCIENCE. *Compiled by Robert L. Weber and Eric Mendoza.* A collection of jokes, short essays and poems on humor and humanity in physics and related sciences. **$12.50**

56270. INTRODUCTION TO GENERAL SYSTEMS THINKING. *Gerald M. Weinberg.* One of America's leading experts on general systems theory provides a non-technical, yet thorough account of this powerful new method for solving complicated problems of a multi-disciplinary nature. **$16.95**

86600. THE WEATHER MACHINE. *Nigel Calder.* One of the leading popular-science writers of our times takes us to the outposts of meteorological research to explain how the weather machine works and why it is changing gear. **$14.95**

NO-RISK GUARANTEE! Examine your 3 introductory books for 10 days. If not delighted, simply return them and your membership will be canceled. You pay nothing, owe nothing!

70240. PRE-COLUMBIAN CITIES. *Jorge E. Hardoy.* Remarkable study of the prehistoric civilizations of the Western hemisphere. Implications for the contemporary urban crisis in America. **$19.95**

86070. WANDERING LANDS AND ANIMALS. *Edwin H. Colbert.* A boldly original epic history of our planet's drifting continents and the spread of animals through the ages. Replete with charts, tables, maps, illustrations and impressive glossary. Recounts the finding of *Lystrosaurus* in Antarctica, now regarded as keystone evidence to explain continental drift. **$12.50**

44941. THE ENCYCLOPEDIA OF PHYSICS. Second Edition. *Edited by Robert M. Besancon.* Updated and expanded edition of a modern classic. Over 300 distinguished authorities offer incisive, authoritative information on every topic in physics today. *Counts as 2 of your 3 books.* **$37.50**

81820. SYMMETRY IN SCIENCE AND ART. *A.V. Shubnikov and V.A. Koptsik.* This new edition of a classic work on symmetry explores every facet of the subject. *Counts as 2 of your 3 books.* **$35.00**

44861. THE ENCYCLOPEDIA OF CHEMISTRY. Third Edition. *Edited by Clifford A. Hampel and Gessner G. Hawley.* Completely revised and greatly expanded new edition of a classic. Over 1200 pages, invaluable charts, diagrams, and illustrations. *Counts as 2 of your 3 books.* **$39.50**

67190. THE PATH TO THE DOUBLE HELIX. *Robert Olby.* A captivating history of twentieth century experimental biochemistry and a highly readable account of the physical and chemical nature of the gene. *Counts as 2 of your 3 books.* **$23.50**

41610. DATA ANALYSIS FOR SCIENTISTS AND ENGINEERS. *Stuart L. Meyer.* Incredibly useful book covering all the statistical methods, concepts and analysis techniques that any experimenter will probably ever need. 513 pages. **$16.95**

How the Club Operates • The Book Club **News**, describing the coming Main Selection and Alternate Selections, will be sent to you 15 times a year at 3-4 week intervals. • If you wish to purchase the Main Selection, do nothing and it will be shipped to you automatically. • If you prefer one of the Alternates, or no book at all, simply indicate your decision on the reply form always enclosed with the **News** and mail it so that we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News**, you should ever receive a Main Selection without having had the 10 day consideration period, that Selection may be returned at Club expense. • Continued membership (after purchasing 3 books) brings you the added benefit of our Bonus Book Plan.

BOOKS

Newton's theology: an effort to study the universe through biblical exegesis

by Philip Morrison

THE RELIGION OF ISAAC NEWTON, by Frank E. Manuel. Oxford University Press ($11.25). Newton sent to the printer, apart from his scientific works, only one volume of any importance: a meticulous textual study of dating in ancient history, his *Chronology of Ancient Kingdoms Amended.* (It was not much different, in an era unable to read hieroglyphic or cuneiform, from the understanding of the authors of classical antiquity.) But he had also produced, and he left behind, bundles of refractory manuscript that were never published, and were hardly even read, before they were scattered by an auction in 1936. The bulk of them have been reassembled by three zealous collectors, a famous Cambridge economist-don, an American market analyst and a Yale Orientalist. There are more than a million words on religion alone. (Many of the papers, to be sure, are different drafts of the same work.) About a third of this brief scholarly book is citation from the manuscripts themselves. It is difficult matter, including both Newton's rules for interpreting Scripture and his substantive glosses on enigmatic Bible text; for example, he works out the volume of the New Jerusalem after Judgment Day.

Professor Manuel is a penetrating (although not a scientist's) biographer of Newton. He has already given us one careful study of Newton as a historian and perhaps the best single evocation of Newton as a person, science apart. Now he has gone deep into the papers held in the earthly Jerusalem, the gift of Professor A. S. Yahuda to the university there. Einstein once praised these papers as a clue to Newton's "spiritual workshop." The great historian of science George Sarton was indifferent to them, considering that they could be of no concern to a scientist. Manuel is surely justified in his sense that Newton was one man, immersed in a century of theological warfare, ruled by a passion for origins and for a simple truth that lay behind the shifting appearances of the world. For Newton and his time there were two ways to examine the universe God had made, one through the book of nature, the other through Scripture. To read either book specialized study was needed, along with acute and learned reasoning. For each Newton set the same great criterion: "Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.... He is the God of order and not of confusion. And therefore as they that would understand the frame of the world must indeavour to reduce their knowledg to all possible simplicity, so it must be in seeking to understand these visions."

How Newton tried! St. John saw frogs issue from the mouth of the dragon and the mouth of the beast and the mouth of the false prophet. For Newton frogs had to mean papal idolaters and their practices. Why so? Newton found in a series of authors—Aristophanes, Aristotle, Ovid, Origen, even Grotius—a broad consensus that frogs were like babblers, impostors, garrulous empty teachers. What else could John have meant but idolaters? And who were the idolaters of Newton's day but the Papists? Once the dates of events were clear to him and the symbol-coding had been decided, a Newton could read out the scattered visions as prophecy and narrative. The correspondence with past history seemed perfect. He never found a prediction as clear and striking, however, as the one Halley could make of the returning comet from nature's mathematical book; the Book of the Apocalypse remains to most of us still no ephemeris but a venerable enigma. The conclusions of great Newton, unread and unheeded, seem self-deluding and awry.

Three points are made here, less germane to a narrowly scientific reader than to a reader concerned with the fuller history of ideas. First, it was not the London civil servant in his dotage who was the tireless theological synthesizer. There is not much doubt that the Cambridge professor who used prism and mirror and who wrote the wonderful *Principia* was the very man who pursued the symbols of the Revelation. The salient pages are roughly datable. While Newton was a young genius ("in the prime of my age for invention") his religion was already personal and dominating, but it had not yet turned exegetical. Second, Newton argued against abstract metaphysics, not only with respect to Leibniz but also within his faith. For him the early Fathers "perplexed the Church with metaphysical opinions." He cared nothing for the subtle substances of the Trinity (although he was no Unitarian, even if he was more sympathetic with that doctrine in private than he ever expressed to the world). His all-powerful God was bound up with plain words, the words of Scripture, and not with philosophical abstractions; Newton reworked texts, but he stuck to concrete language. Neo-Platonism, Deism and indeed much Enlightenment metaphysics were nourished by Newtonian thought, perhaps, but never by Newton. Third, we are offered evidence (there can hardly be proof) that this complex and withdrawn, if also litigious, man was inwardly moved by a quest for a Father in Heaven. Newton, born after his father's death—and significantly on Christmas Day—could have no earthly father to recall. Writes Manuel: "To believe that one had penetrated the ultimate secrets of God's universe and to doubt it, to be the Messiah and to wonder about one's anointedness, is the fate of prophets. Newton's conviction that he was a chosen one of God, miraculously preserved, was accompanied by the terror that he would be found unworthy and would provoke the wrath of God his Father. This made one of the great geniuses of the world also one of its great sufferers."

Newton's manuscripts have now been conned for their mathematics, their every implication for the larger books, their metaphysics and their theology. One phase of his unending industry remains unexamined in detail, since it somewhat passes the bounds of pen-and-pencil scholarship. That is his chemistry, a strange legacy of alchemical writings that remains a challenge to one who will

enter it someday, no less learned in amalgams than in Hermes Trismegistus. We await assayers to come.

THE PHYSICS OF TIME ASYMMETRY, by P. C. W. Davies. University of California Press ($15.75). A rain of contentious and technical papers has washed over the physics of time asymmetry, eroding away any means of general access. "Great confusion and misunderstanding" surround the subject; learned authors seem almost to avoid common ground. We owe a real debt to this work by a young British cosmologist: a clarifying, brief and not very mathematical volume that makes a single, sensibly balanced topic out of the varied approaches—with perhaps an agreeable bias toward cosmology.

The title itself is material. It is not time that has an arrow, a direction, a flow, an asymmetry, whatever the happy phrase may say. ("In an empty universe such things do not exist.") It is the world's contents—matter and radiation and their changes—that display a structure distinguishing past from future. More philosophical issues such as the subjective status of the "now" are deliberately set aside. Indeed, the first chapter concludes with a view of time itself as an entirely passive marker, a line in space-time that special relativity has taught us how to label for each observer.

Thermodynamics and statistical mechanics are then discussed in their relevant aspects, with the right equations very much reduced to their formal essentials and most of the comment in ordinary language. In this context the first major answer is reached, which we owe to the philosopher Hans Reichenbach. Our thermodynamic experiments refer, he showed, to partially ordered systems that are part of the main environment only until we make some change that isolates them for a time. The ice cube melts in the glass once we pour the drink. Before that the system ice plus drink never existed. Within the glass melting occurs, whatever the universe at large thinks of it. Asymmetry is unambiguous in the huge majority of such branch systems. They were marked with low entropy by birth, so that they display the growth of disorder with time. It is of no use to argue that the branch's entropy could also have been high before the time of separation, because the branch system has no past at all.

The second great point is the role of approximate macroscopic description, technically called coarse-graining. The ice cube is what melts; its constituent molecules are not described one by one either before or after the event. Since Gibbs, at least, we have known that such a description leads to asymmetry by throwing away the subtle correlations the actual dynamical past imposes almost unobservably. The critics maintain that this is a mere approximation, too subjective to credit. But the third point is the rejoinder: The inordinately weak coupling of every "isolated" system with the outside world would in the end mix up even the inherent fine-grained memories of motion. Of this Davies is unconvinced. The problem deserves further study, but this view seems to fit the flow of his argument.

We need big, low-entropy systems to feed our small, active branch systems. Locally the trunk is the sun, the fount of all our free energy. Only gravitation can supply such a fountain, the author maintains. As gravitating systems collapse they naturally divide into two concentric parts: a dense space-ordered core of low entropy and a diffuse fringe of high-entropy debris. In this way the eventual origin of all thermodynamic irreversibility is traced to gravity. Here one may demur a little. The gravitational phenomenon, although it is real enough, does not seem all that different from the behavior of real molecular gases, which also condense into two density phases when they are cooled. The big world is gravity-ruled, all right, but even if it were not, asymmetry might well arise out of the other interactions.

Davies' account of cosmology, with its distinction between order-preserving universal expansion and disorderly starlight, is fresh. He also treats such fashionable related issues as Olbers' paradox and black holes. The next two chapters expound on somewhat more restricted questions. Two distinct sources of asymmetry arise in the time-alternate solutions of the Maxwell equations (advanced or retarded by transit time) and in quantum mechanics. Here the general reader must enter cautiously; the questions are both technical and, like the tachyon and the problem of measurement in quantum mechanics, far from settled in principle.

The book closes with a chapter, built on less than hard rock, that treats the newer ideas of cosmology, from the steady state to universal oscillations to closed time. Davies' own model—a two-step universe, one forward, one back—is a neat realization of the loop of closed time.

A study of this book and its valuable references will go a long way toward ordering a field itself rather disordered by time. One may remain a bit skeptical of the closure of the subject, pretty well hinted at here. Would the life of experimenters in a well-furnished cosmic bomb shelter be so different from our own life, extragalactic astronomy aside? Time's arrow may turn out to have a hierarchy of origins; their order of importance will remain a matter for later decision.

BRASSEY'S INFANTRY WEAPONS OF THE WORLD: 1975, edited by J. I. H. Owen. Hippocrene Books, Inc. ($49.50). Power flows more clearly these days out of the barrel of a gun held by a man who lives right here on the ground than from multiple jets flown overhead from distant bases. This thick volume, with extra-big halftones, wide margins, crisp technical descriptions and knowing, gossipy little essays on history and rationale, describes the gun the man on the ground holds, whoever he is. The last pages list the armies from Abu Dhabi to Zaïre and tabulate their standard equipment in the categories of pistols and revolvers; submachine guns, rifles and carbines; machine guns, mortars, antitank missiles and rockets. The preceding chapters have described and have almost always illustrated the devices named plus a good many more lethal weapons a man might carry: flamethrowers, grenades, mines, mine detectors and novel accessories such as infrared viewers and battlefield radar. The major categories are treated comprehensively and the minor ones are widely sampled. The entire presentation stems from the familiar London community of worldwide arms intelligence, the source of most of the authoritative public information of this kind since before 1914. The editor, who has several expert compilers in support, is listed as "Major General, OBE late Royal Marines."

Here is the famous Kalashnikov assault rifle, AK47 and AKM models, the newer AKM lighter by a kilogram, using more cheap steel stampings and yet much the same. This weapon is currently produced as the principal arm of the infantryman in most Warsaw Pact countries, the Chinese People's Republic, Finland and Yugoslavia. Today the gun bulks as large economically as some jet-fighter models; its total production run can be valued well into the billion-dollar range. Thirty million or more have been made: all gas-operated automatic rifles, holding thirty .30-caliber (metric 7.62-millimeter) rounds in a curving magazine, fired at 600 rounds per minute out to an effective range of 400 meters. Submachine guns are here, from the famous Thompson of 1921 (accurate, but heavy and expensive) to the very cheap, light

Sten gun of 1940 British design (more than four million were made in World War II) and the U.S. "grease gun." ("Today's weapons are basically only refinements of the designs laid down 30 or 40 years ago.") The arms race is slower here, although real enough. Copies of the enemy's designs are commonplace (the grease gun is made in China too) and the ability to use his ammunition is held to be advantageous. Reliability and ease of production are important criteria; performance is more nearly equal among similar weapons. One nine-inch (stock folded), three-and-a-half-pound Ingram submachine gun, the newest U.S. model, fires 9-mm. shorts at 20 rounds per second from a magazine and is appraised as the "ideal personal weapon for clandestine operations." (It looks like a stapler with a handle.)

Here is the newest class of weapons too: the suitcase-weight, low-level surface-to-air heat-seeking missile. The Russian version is called Strela ("arrow"), the U.S. has Red Eye and the U.K. has Blowpipe. In all of these a hand-held launch tube can eject a rocket to fly free at Mach 1.5 for three or four kilometers, seeking the infrared glow of an aircraft-engine exhaust. Current models have had success against attack aircraft and helicopters. Faster, larger aircraft have so far been less endangered, but new models—the U.S. Stinger for one—are in train that are said to have speed and guidance good enough to threaten the costly Mach 2 jets.

The panoply of Vietnam is described. The AK47, the 120-mm. Stokes mortar (a five-or-six-man weapon of Russian design and Chinese make) and the double-tube rocket launcher (its 20-kilogram high-explosive spin-stabilized missiles reaching out eight kilometers, not very accurately) were all well known to newspaper readers, at least by name. The U.S. fielded counterparts to all of them. Most of the weapons are assessed by the expert reviewers. Very practical properties are esteemed in infantry combat. The new Beretta submachine gun (Italy) is marked by deep grooves that "allow for a large amount of dirt and foreign matter to be absorbed" before the gun malfunctions. The key to the air-cooled machine gun, now ubiquitous, was the easily detachable barrel. The first true general-purpose machine gun (Mauserwerke AG, first series production in 1936) was issued with an asbestos glove to handle that particular chestnut.

Mines are a gloomy art. Many are aimed against man, with cheap and cunning ways of jumping up into the air to spread their cruel designed-in fragments

more widely. TNT blocks housed in the two pockets of a canvas jacket worn by trained Russian dogs were much used against German tanks. ("Even today there is still said to be a shortage of dogs in the U.S.S.R.") The definitive undetectable mine is a U.S. innovation that is not yet in service. A special liquid explosive is poured or sprayed into the topsoil. A small detonator ignites the charge to scatter the earth itself. Liquid Astrolite, like certain declarations of state, becomes inoperative after four days. Proliferation of all these arms is wide; surplus stocks diffuse in bulk, and manufacturers are listed here from 30 countries.

COWS, PIGS, WARS AND WITCHES: THE RIDDLES OF CULTURE, by Marvin Harris. Random House ($7.95). Ambitious, argumentative, acute and delightful, this brief, informal book is a connected series of case studies seeking the rationale of cultural practices over the widest range of societies. The 1,000 pig-loving New Guinea mountain gardeners with their elaborate ritual-warfare cycle and the war-torn Europe of Huss and Cromwell, burning its witches, are here alike. Professor Harris, a distinguished Columbian anthropologist, does not rest on the superficialities of a personal view of human nature and its inner light and darkness. For him the world's cultures, small and large, are evolving structures that have become more or less adaptive in their social and economic settings. They make sense broadly, although that sense must take account of their fundamental divisions into distinct social classes and their needs for self-maintenance.

Ruth Benedict wrote on the potlatch as a bizarre search for status and concluded that the entire economic system of the Indians of the Northwest was "bent to the service of this obsession." Harris turns the argument around: for him status rivalry, although it was real enough, was "bent to the service of the economic system." There were surpluses in that rich region, but the luck of the fish run and the fruit harvest fluctuated from village to village. Each year "the haves gave and the have-nots took." This is an adaptive form of pooling, an incentive to production, one step on the way to a society with a defined ruling class and then to the state itself, able to build pyramids and to dig a thousand missile silos. The Bushman and the Eskimo, egalitarians and hunters, understand that "gifts make slaves just as whips make dogs." Among the Kaoka in the poor Solomons there are self-elected "big men"; they offer a frugal equivalent of potlatches—for which they must work hard, consume less and worry more than anyone else. Prestige is their only reward. Among the richer Kwakiutl the same notion spawned hereditary managerial chiefs.

Egalitarians reject that entire road. Among the Malayan Semai "it is rude to be openly grateful." The Kalahari Bushmen denigrated as a "worthless animal" the gift ox their Toronto guest offered them in feast. "We refuse one who boasts," they said, "for someday his pride will make him kill somebody." Writ large, the fear was sound. Societies that developed gifts went on to governing classes, the drive to hard work, maximal production and large populations. They did in time destroy or displace simpler societies. Reciprocity gave way to the potlatch, the potlatch to the king, the king to the capitalist and the Five-Year Plan. The dependable hunter never boasts, never hints he is making a gift as he cuts up and divides the game; the hunting way of life cannot sustain widespread driving hard work and consequent economic growth. "If he got his followers to work like the Kaoka for a month, an aspiring Bushman big man would kill or scare off every game animal for miles around and starve his people to death before the end of the year." Human cultures make sense in their own terms, and the individual nature of the unconscious, the "drives to aggression" and to prestige, conform to the human way rather than determine it.

War, Christianity, cargo cults, witch-burning and near-universal (but not intrinsic) male supremacy are given discussions no less audacious. For Harris war is adaptive when it is population-controlling; male leadership is adaptive when hand-held weapons are the basis of warfare. Neither necessity is present today; sexual equality, we must hope, will arise from contraception and the elimination of police forces and armies rather than out of the decimation of peoples and the perfection of weapons that place no premium on large muscles. The greatest sacred cow is not that scrawny Hindu animal, source of fuel and tractive power for the poor man, scrounging for its fodder. "If you want to see a real sacred cow, go out and look at the family car." It is not an excess of objectivity we suffer under today; the "new consciousness" will bring witches and messiahs as always. The witch has returned because such an institution "blunts and befuddles the forces of dissent" as it did in the 15th century. "You can't make a revolution if everybody does his own thing. To make a revolution, everybody must do the same thing."

This is an unusually provocative work. Of course there are objections, plenty of them. The material and the moral worlds are strongly coupled. Maybe edible pigs are truly better for poor tropical farmers than uneaten cattle, as China is agriculturally better off than India. Analysis along these lines is surely the best way to find out. To start discussion by reading these lively chapters is the best way of "doing the same thing" that one has seen lately. (The publisher goofed by not providing an index.)

PEST CONTROL: A SURVEY, by Arthur Woods. Halsted Press Division of John Wiley and Sons ($29.50). Your commercial greenhouse is lined with tender young cucumber vines. Bringing them to fruit is a season's task, demanding painstaking care. A heavy investment is at stake. The key pest of the cucumber is the red spider mite, wretched tiny things likely to concentrate near the tender tops of the plants, where "a comparatively small number can cause severe damage." Yet the first step in modern control, shortly after planting, is to release "ten to twenty red spider mites ... onto each plant." This heroic inoculation has its rationale: 10 days later you set out two specific allied predatory mites on every other plant. If the predators find enough red-spider-mite prey to establish themselves, they will protect your crop. The method is promising: the yield goes up 25 percent over chemical control, since the usual sprays weaken the plants, and you save the labor costs of spraying. The task of persuading growers to start the season by infesting their own greenhouses with the major pest is no small problem. Yet it looks as though wide application of this scheme is not far off in the British greenhouse-crop trade.

This thorough overview of the ways to protect growing crops from attack by other organisms is the work of an Australian entomologist with wide experience and a mathematical taste. It is really a broad introduction to applied ecology. A discussion of the general importance and economics of pest control is followed by a chapter on "the numbers of animals," a summary at a simple level of the description and measurement of animal populations. The subsequent chapters give a fairly technical but not detailed survey of the main means of control. First come chemical pesticides of all kinds, including herbicides; then comes biological control by insects, mites, vertebrate animals and

microorganisms. Next follow the ingenious "autocidal" schemes, in which a pest species is manipulated to destroy its own kind, say by the release of enough sterile screwfly males to ensure that most matings by wild females prove fruitless. Attractants of various kinds, including specific insect sex pheromones, are treated, as are the classical schemes: resistant varieties, crop rotation, fences and barriers—and flyswatters. The last chapter makes the book's point. Integrated control is the right way, making use of any and all means as called for in a rational campaign based on knowledge of the agro-ecosystem and the population dynamics of the forms concerned.

The chemicals such as DDT were so cheap and at first so effective that they became hazards themselves as their use became indiscriminate and excessive. The traditional means of control had been empirically integrated, but "this art was largely forgotten during the 'dark ages' of DDT." Yet DDT cannot be entirely dismissed; its unmatched persistence on mud walls and its low price make it the first choice still in countries where mosquitoes that carry malaria are its target. The World Health Organization considers that DDT will remain essential. Its side effects extract a price; the selection of resistant strains of the target forms remains its limitation.

Wheat-stem rust is fought on the Great Plains in the traditional way, by meticulous eradication of its intermediate host, the barberry shrub. A single gene confers rust resistance on wheat, and the plant breeders steadily produce new resistant strains of the cereal. By the same token a single mutation in the pathogen can produce a new race of fungus able to attack any given new variety. Here is the dynamical problem plain; it has no static solution. Eternal vigilance is the price of yield.

Dr. Woods, although he offers little support for his view, is an enthusiast for computer simulation. Such schemes are attractive to people who face many choices, but they will prove to be no panacea. The entire burden of this judicious work is the complex nature of real ecosystems, which easily outreach our models and clearly plead for modest step-by-step action watching for feedback. Indeed, the intricate ecology of petrochemical firms, their eager salesmen and their weevil-beset customers may be the best clue to the mistaken simplicities of the era of chlorinated hydrocarbons in plenty. The more they sprayed, the worse the crop. But who could have programmed the chemical companies?

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the list below helpful.

INTERNATIONAL COMPARISONS OF MEDICAL CARE

STATISTICAL INDICATORS FOR THE PLANNING AND EVALUATION OF PUBLIC HEALTH PROGRAMMES. Fourteenth Expert Committee on Health Statistics, World Health Organization Technical Report Series No. 472. World Health Organization, 1971.

HEALTH CARE: CAN THERE BE EQUITY? THE UNITED STATES, SWEDEN AND ENGLAND. Odin W. Anderson. Wiley-Interscience, 1972.

INTERNATIONAL MEDICAL CARE. Edited by John Fry and W. A. J. Farndale. Medical and Technical Publishing Co., Ltd., 1972.

HEALTH CARE: AN INTERNATIONAL STUDY. Edited by R. Kohn and K. L. White. Oxford University Press, 1975.

GIANT RADIO GALAXIES

THE EVOLUTION OF RADIO ASTRONOMY. J. S. Hey. Elek Science, London, 1972.

ACTIVE GALAXIES WITH RADIO TRAILS IN CLUSTERS. G. K. Miley, G. C. Perola, P. C. van der Kruit and H. van der Laan in *Nature*, Vol. 237, No. 5353, pages 269–272; June 2, 1972.

RADIO GALAXIES AND QUASARS. K. I. Kellermann in *Galactic and Extragalactic Radio Astronomy*, edited by G. L. Verschuur and K. I. Kellermann. Springer-Verlag, 1974.

3C 236, DA 240: THE LARGEST RADIO SOURCES KNOWN. A. G. Willis, R. G. Strom and A. S. Wilson in *Nature*, Vol. 250, No. 5468, pages 625–630; August 23, 1974.

HOW BACTERIA SWIM

CHEMORECEPTORS IN BACTERIA. Julius Adler in *Science*, Vol. 166, No. 3913, pages 1588–1597; December 26, 1969.

HOW TO TRACK BACTERIA. Howard C. Berg in *The Review of Scientific Instruments*, Vol. 42, No. 6, pages 868–871; June, 1971.

CHEMOTAXIS IN *ESCHERICHIA COLI* ANALYSED BY THREE-DIMENSIONAL TRACKING. Howard C. Berg and Douglas A. Brown in *Nature*, Vol. 239, No. 5374, pages 500–504; October 27, 1972.

BACTERIA SWIM BY ROTATING THEIR FLAGELLAR FILAMENTS. Howard C. Berg and Robert A. Anderson in *Nature*, Vol. 245, No. 5425, pages 380–382; October 19, 1973.

CHANGE IN DIRECTION OF FLAGELLAR ROTATION IS THE BASIS OF THE CHEMOTACTIC RESPONSE IN *ESCHERICHIA COLI*. Steven H. Larsen, Robert W. Reader, Edward N. Kort, Wung-Wai Tso and Julius Adler in *Nature*, Vol. 249, No. 5452, pages 74–77; May 3, 1974.

FLAGELLAR ROTATION AND THE MECHANISM OF BACTERIAL MOTILITY. Michael Silverman and Melvin Simon in *Nature*, Vol. 249, No. 5452, pages 73–74; May 3, 1974.

THE CAUSES OF BIOLOGICAL DIVERSITY

GENETIC DIVERSITY AND NATURAL SELECTION. J. J. Murray. Hafner Press, 1972.

THE GENETIC BASIS OF EVOLUTIONARY CHANGE. R. C. Lewontin. Columbia University Press, 1974.

NATURAL SELECTION AND HEREDITY. Philip M. Sheppard. Hutchinson University Press, 1975.

THE PERCEPTION OF SURFACE COLOR

AN INTRODUCTION TO COLOR. Ralph M. Evans. Wiley-Interscience, 1948.

THE PERCEPTION OF BRIGHTNESS AND DARKNESS. L. Hurvich and D. Jameson. Allyn & Bacon, Inc., 1966.

SURFACE COLOR PERCEPTION. Jacob Beck. Cornell University Press, 1972.

THE FLOOR OF THE MID-ATLANTIC RIFT

UNDERSTANDING THE MID-ATLANTIC RIDGE: A COMPREHENSIVE PROGRAM. Edited by J. R. Heirtzler. National Academy of Sciences, 1972.

FLOW OF LAVA INTO THE SEA. James G. Moore, R. L. Phillips, Richard W. Grigg, Donald W. Peterson and Donald A. Swanson in *Geological Society of America Bulletin*, Vol. 84, No. 2, pages 537–546; February, 1973.

FAMOUS: A PLATE TECTONIC STUDY OF THE GENESIS OF THE LITHOSPHERE. J. R. Heirtzler and X. Le Pichon in *Geology*, Vol. 2, No. 6, pages 273–274; 1974.

ATLANTIC OCEAN FLOOR: GEOCHEMISTRY AND PETROLOGY OF BASALTS FROM LEGS 2 AND 3 OF THE DEEP SEA DRILLING PROJECT. Fred A. Frey, Wilfred B. Bryan and Geoffrey Thompson in *Journal of Geophysical Research*, Vol. 79, No. 35, pages 5507–5527; December 10, 1974.

A PRE-COLUMBIAN URBAN CENTER ON THE MISSISSIPPI

ARCHAEOLOGICAL PHASES AT CAHOKIA. Melvin L. Fowler. Papers in Anthropology No. 1, Illinois State Museum Research Series, 1972.

EXPLORATIONS INTO CAHOKIA ARCHAEOLOGY. Edited by Melvin L. Fowler. Bulletin No. 7, Illinois Archaeological Survey, Inc., 1973.

CAHOKIA: ANCIENT CAPITOL OF THE MIDWEST. Melvin L. Fowler. Addison-Wesley Module in Anthropology No. 48, 1974.

THE STELLAR-ORIENTATION SYSTEM OF A MIGRATORY BIRD

DIE STERNENORIENTIERUNG NÄCHTLICH ZIEHENDER GRASMÜCKEN. E. G. Franz Sauer in *Zeitschrift für Tierpsychologie*, Vol. 14, pages 29–70; 1957.

BOBOLINK MIGRATORY PATHWAYS AND THEIR EXPERIMENTAL ANALYSIS UNDER NIGHT SKIES. William J. Hamilton III in *The Auk*, Vol. 79, pages 208–233; 1962.

MIGRATORY ORIENTATION IN THE INDIGO BUNTING, PASSERINA CYANEA. PART II: MECHANISM OF CELESTIAL ORIENTATION. Stephen T. Emlen in *The Auk*, Vol. 84, pages 463–489; 1967.

BIRD MIGRATION: INFLUENCE OF PHYSIOLOGICAL STATE UPON CELESTIAL ORIENTATION. Stephen T. Emlen in *Science*, Vol. 165, No. 3894, pages 716–718; August 15, 1969.

CELESTIAL ROTATION: ITS IMPORTANCE IN THE DEVELOPMENT OF MIGRATORY ORIENTATION. Stephen T. Emlen in *Science*, Vol. 170, No. 3963, pages 1198–1201; December 11, 1970.

BIRD MIGRATION. Donald R. Griffin. Dover Publications, Inc., 1974.

MATHEMATICAL GAMES

THE MATHEMATICAL IMPLICATIONS OF ESCHER'S PRINTS. H. S. M. Coxeter in *The World of M. C. Escher*, edited by J. L. Locher. Harry N. Abrams, 1971.

SYMMETRY IN SCIENCE AND ART. A. V. Shubnikov and V. A. Koptsik. Plenum Press, 1974.

THE AMATEUR SCIENTIST

SAILING THEORY AND PRACTICE. C. A. Marchaj. Dodd, Mead & Co., 1964.

INTRODUCTION TO PLANETARY PHYSICS. William M. Kaula. John Wiley & Sons, Inc., 1968.

The right way to use an aluminum can is to use it again.

The aluminum can is lightweight, chills quickly and keeps beverages fresh and flavorful. It's also easy to recycle. Used aluminum cans are baled or shredded and shipped out from the reclamation center, then remelted and fabricated into brand-new cans. Recycling is important because it saves energy and gives aluminum new life, over and over. That's why Alcoa supports recycling programs throughout the country.

Aluminum cans are being remelted with only five percent of the energy it took to make molten metal from the ore in the first place. Recycling saves natural resources, too. You can see that recycling cans is beneficial to all of us. In the future, the high value of aluminum scrap could help make it economically attractive to recycle municipal solid waste.

The point is that a used aluminum can has become far too valuable to waste.

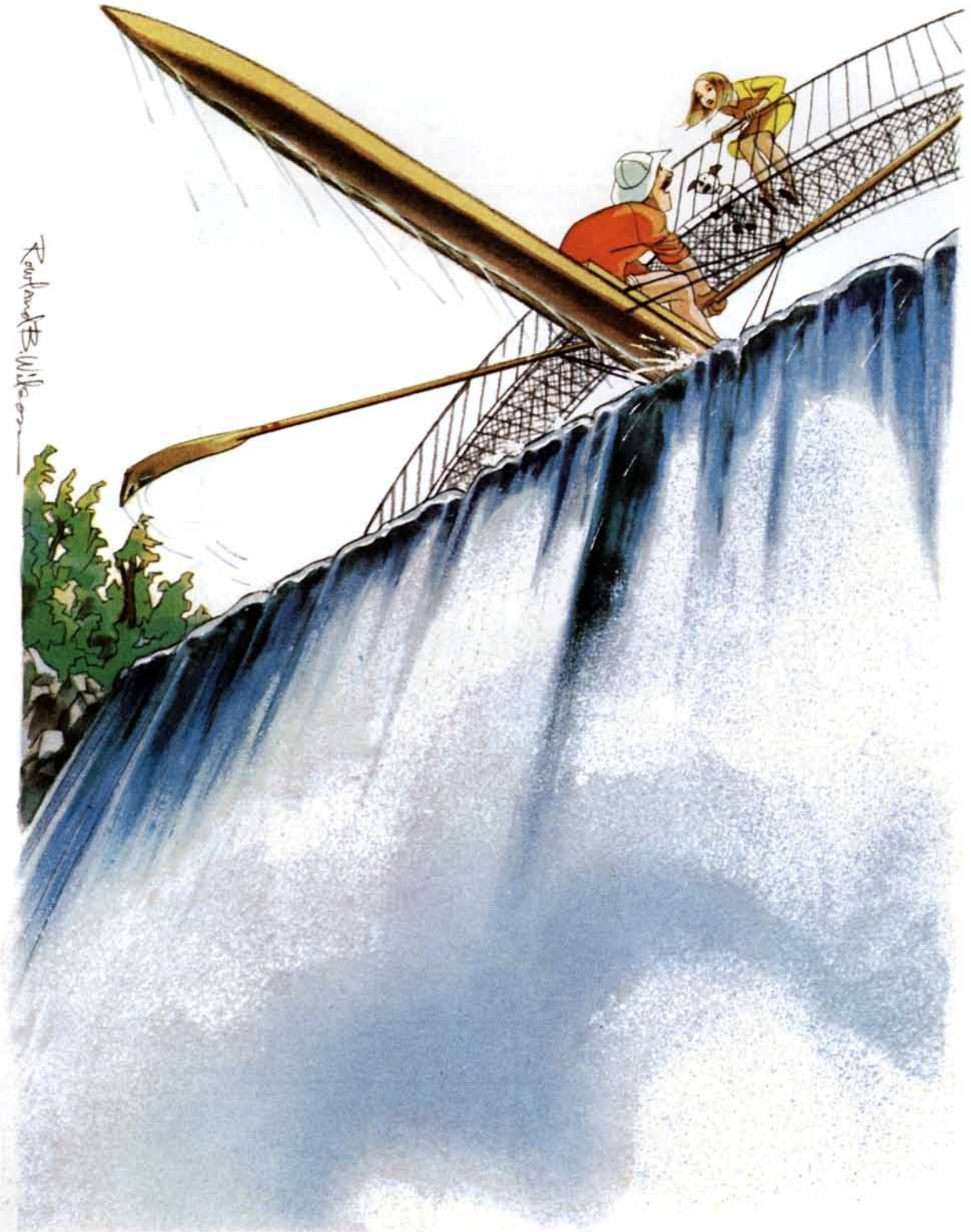
Alcoa's "Yes We Can" recovery program is just one way of making certain that we don't ignore one of the valuable resources of our country.

Over 30 million dollars have been paid to collectors through the combined recycling efforts of the aluminum and beverage industries. Since 1970, almost six billion cans have been recovered. There is no doubt that the collection and recycling of aluminum cans is working. More than two billion were collected last year, and the number is growing.

For more information on aluminum can recycling and how one community established a reclamation program, please write for our free brochures. Aluminum Company of America, 986-H Alcoa Building, Pittsburgh, PA 15219.

The reasons for using aluminum are found in aluminum itself.

ALCOA



"My insurance company? New England Life, of course. Why?"

Our crew also offers mutual funds, variable annuities and investment counseling.