SCIENTIFIC AMERICAN

ELECTRONICALLY STEERED RADAR

$2.50

February 1985

The Mercedes-Benz 300D Sedan, 300TD Station Wagon and 300CD Coupe: with their Turbodiesel performance, they are

The Mercedes-Benz Turbodiesels for 1985: still the most powerful line of diesels sold in America.

THE MERCEDES-BENZ 300D Sedan, 300TD Station Wagon and 300CD Coupe represent three variations on a radical theme: the idea that dramatic over-the-road performance can be blended with diesel efficiency and stamina.

The idea works. These Mercedes-Benz Turbodiesels *move*. With accelerative energy and cruising ease worthy of gasoline-powered cars. With power enough to flatten hills and make quick work of sudden passing maneuvers.

TURBODIESEL POWER, DIESEL DURABILITY

Yet consider the bottom line. The Turbodiesel you will be living with and maintaining and paying the bills for, year in and year out, is a true-blue diesel. No complex electrical system. No conventional tune-ups. A durability factor that has become part of automotive folklore.

The key to the Mercedes-Benz Turbodiesels' performance is less the *turbo* than the *diesel*–its three-liter, five-cylinder engine.

It is unique, a high torque powerhouse so advanced that it even oil-cools its own pistons as they move.

Turbocharging any engine boosts its power. Turbocharging this engine boosts its power–by 42 percent in models sold on the West Coast, by 45 percent in models sold elsewhere.

Many makers have aped the Turbodiesel idea since Mercedes-Benz pioneered it in production automobiles in 1978. Scant surprise that no maker has yet aped the Mercedes-Benz Turbodiesels' vivid level of performance.

The Turbodiesels rank not only as the most powerful but also the most *varied* line of diesels sold in North America today.

SEDAN, STATION WAGON AND COUPE

The four-door 300D Sedan accommodates five persons and a gaping 12-cu.-ft. trunk within a wheelbase of just 110 inches, helping lend near sports-sedan agility to this familiy-sized automobile.

"The 300D's success in striking a balance between ride comfort and handling response," reports one automotive journal, "is equalled by less than a handful of other cars in the world."

The 300TD Station Wagon interlaces the driving pleasures of a Mercedes-Benz with the workhorse utility of a five-door carry-all. Total cargo capacity well exceeds 100 cu. ft. A hydropneumatic *leveling* system is integrated with the rear suspension, to help keep the vehicle riding on an even keel–whether the load is heavy or light.

EXOTIC, YET PRACTICAL

The 300CD Coupe is the world's only limited-production two-plus-two diesel touring machine. It sits on a taut 106.7-inch wheelbase–one secret of its quick-witted

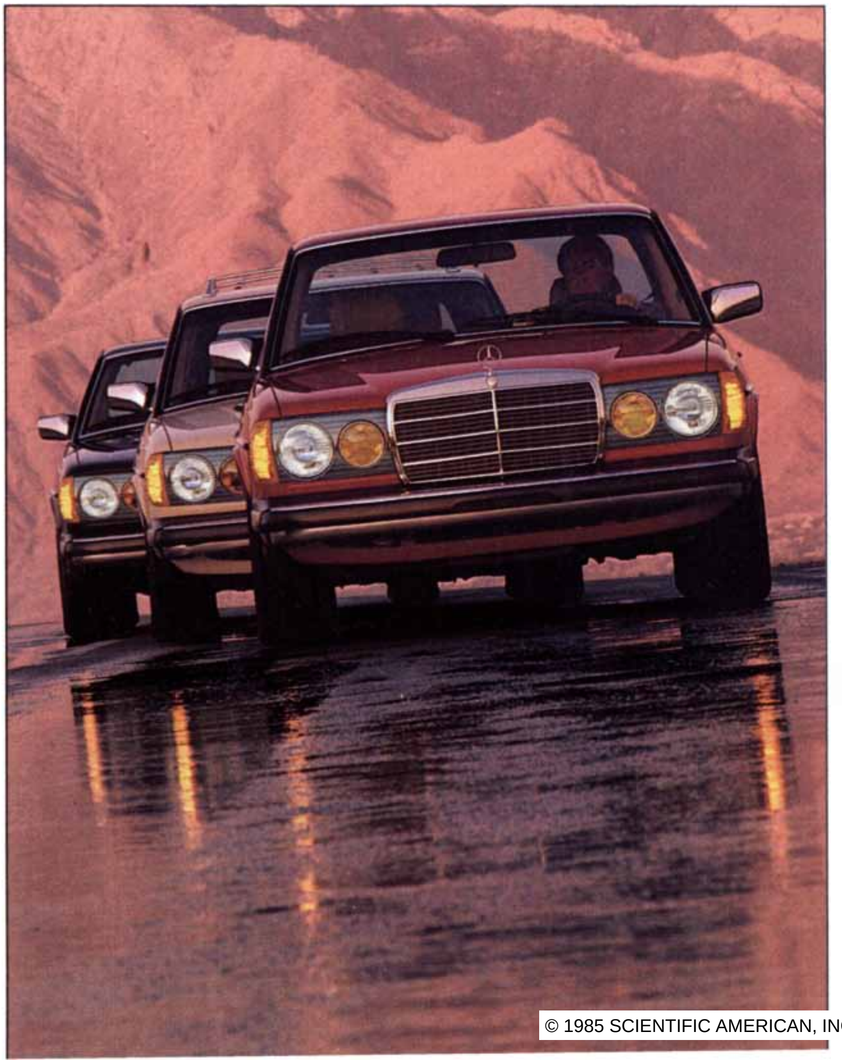
1985 TURBODIESELS

diesels apart. With their handling agility and riding comfort and obsessively fine workmanship, they are automobiles apart.

agility. Its graceful coupe body-work, sans central door pillars, is formed in a process involving intensive handworkmanship. The 300CD is that rarity of rarities, an automobile both highly exotic *and* relentlessly practical.

Sedan or Station Wagon or Coupe, Mercedes-Benz Turbodiesel power is harmonized with high standards of performance in every sense of the word.

From suspension to steering to brakes, every Turbodiesel is engineered to be a precision driving instrument. "There's a cornucopia of driving delights at your disposal," concludes *Car and Driver*–suggesting that in driving precision there is driving pleasure.



From biomechanically correct seats, to a superb automatic climate control system, to the dulling of the outside wind noise to an almost inaudible murmur, remarkable comfort prevails. Virtually every *useful* driving amenity is standard, including an uncannily precise electronic cruise-control unit.

Safety precautions are remarkably comprehensive–both in helping avoid trouble, and in protecting the occupants should trouble occur.

MORE THAN POWER

Ultimately, the Turbodiesels' appeal extends beyond their performance and driving pleasure. There is no more powerful line of diesels sold in North America–and there may be no more versatile, more competent, more timely line of automobiles. In North America, or the world.

Engineered like no other car in the world

SEE YOUR AUTHORIZED MERCEDES-BENZ DEALER


Fly First Class.
Wild Turkey. It's not the best because it's expensive.
It's expensive because it's the best.
Now you can send a gift of Wild Turkey® /101 Proof anywhere* by phone through Nationwide Gift Liquor. Call Toll Free 1-800-CHEER-UP
(Arizona 602-957-4923). *Except where prohibited. Major credit cards accepted. Austin, Nichols Distilling Co., Lawrenceburg, KY © 1985.
Austin Nichols
WILD
TURKEY
KENTUCKY STRAIGHT BOURBON
WHISKEY

Established 1845 **SCIENTIFIC AMERICAN** February 1985 Volume 252 Number 2

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), 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: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

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

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

THE COVER

The cover painting offers a perspective across one of the two 102-foot-wide faces of the PAVE PAWS, an electronically steered early-warning radar built by the Raytheon Company. The PAVE PAWS is an example of a radar technology known as the phased array (see "Phased-Array Radars," by Eli Brookner, page 94). Individual microwave signals from the 1,792 active antenna elements on the 5,354-element array face interfere constructively to produce the beam. Although the beam can be directed as much as 60 degrees away from the array's perpendicular axis, the antenna face is immobile. Instead minute delays, individually calculated, are introduced into the separate signals radiating from each element. The delays shift the phases of the signals in relation to one another so that they interfere constructively at a specified angle with respect to the array's perpendicular axis. Each eight-inch-high element consists of crossed dipoles. These are angled downward to alter electromagnetic interactions of adjacent elements that at certain steering angles can prevent energy from radiating. The metal stubs interspersed among the antenna elements aid in producing a circularly polarized beam, a characteristic that helps to maximize the system's sensitivity.

THE ILLUSTRATIONS

Cover painting by Ian Worpole

Page	Source
12	SCIENTIFIC AMERICAN
18–21	Edward Bell
23	Hillel Burger, Harvard University
33	W. K. Hartmann
36	U.S. Geological Survey, Menlo Park, California
38	Ian Worpole
39	Kerry E. Sieh, California Institute of Technology
40	Andrew Tomko
41–42	Joe Le Monnier
43	Andrew Tomko
45	Steven T. Brentano, University of Iowa (*top*); Klaus M. Esser, Walter Reed Army Institute of Research (*bottom*)
46	Edgar D. Rowton, Walter Reed Army Institute of Research (*top*); Patricia J. Wynne (*bottom*)
47	Laurence Tetley and Keith Vickerman, University of Glasgow
48–49	Ilil Arbel
50	Don. C. Wiley, Harvard University
51	Ilil Arbel
53	Hank Iken, Walken Graphics
54	Hank Iken, Walken Graphics (*top*); H. Pedersen, ESO/La Silla Observatory (*bottom*)
55–57	Hank Iken, Walken Graphics
58	Harvard College Observatory
71–84	George V. Kelvin
87	T. H. Clutton-Brock, University of Cambridge
88	Andrew Tomko
89–92	Patricia J. Wynne
95	Eli Brookner (*top*); Raytheon Company (*bottom*)
96–102	Jerome Kuhl
104–105	Tom Prentiss
106	Quesada/Burke
107–109	Tom Prentiss
113	Quesada/Burke (*top and middle*); Rosemarie Koch (*bottom*)
114	Michael Goodman
115	Ilil Arbel
118	Oleg D. Sherby and Jeffrey Wadsworth
123–126	Michael Goodman

In the photographs you see here, two cars were travelling 50 m.p.h. Over a test track slippery as ice.

The only difference was, one car was equipped with a remarkable new anti-lock brake system, developed by Alfred Teves, a subsidiary of ITT.

It, obviously, wasn't the car on top. When that driver applied his brakes, the wheels locked.

The car began to skid—and the driver lost control.

When the driver in the bottom car applied his brakes, a built-in computer sensed the road beneath the wheels was slippery.

It automatically "pumped" the brakes. Faster than any human driver could.

There was no skidding. No loss of control.

In fact, even if the car were being steered around a curve, it could be braked without skidding.

This unique ITT Teves brake control system will begin appearing in cars this year.

Meantime, it's a good example of the kinds of business activities that ITT is concentrating on these days.

They're businesses where we can apply new technology in imaginative and promising directions—and at top speed.

The best ideas are the ideas that help people. ITT

LETTERS

To the Editors:

Bethe, Garwin, Gottfried and Kendall ("Space-based Ballistic-Missile Defense," by Hans A. Bethe, Richard L. Garwin, Kurt Gottfried and Henry W. Kendall; SCIENTIFIC AMERICAN, October, 1984) have presented an interesting and comprehensive overview of the current state of the technological problems that must be overcome if President Reagan's Strategic Defense Initiative is to become a reality. In the realm of strategic policy, however, their arguments supporting the assertion that strategic ICBM defense is an unwise policy choice are rather less comprehensive and cogent.

One concern stated by the authors is that a U.S. defense that could not fend off a full-scale strategic attack but might be quite effective against a weak retaliatory strike would be particularly provocative to the U.S.S.R. In my view, a more urgent concern is that one day the Soviets themselves may have a defense that could be very effective against a weak retaliatory blow made by the U.S. after a Soviet preemptive attack. There are ample indications that the Soviets are vigorously pursuing their own strategic-defense projects. It is essential that we concern ourselves with the consequences of assuming the Soviets will not proceed with strategic defense if we do not. It is worth noting in this context the staggering efforts the Soviets are undertaking in modernizing their defenses against the air-breathing portion of our strategic triad. The Soviet air-defense forces have already deployed more than 10,000 surface-to-air missile launchers at more than 1,000 sites throughout the Soviet Union.

Indeed, based on observation of Russian policies over the past four decades it can be inferred that a long-range goal of Soviet strategic policy is the attainment of world hegemony. A worrisome consequence of this inference is that the Soviets may not be very interested in equal treaties leading to mutual disarmament. The U.S.S.R. currently enjoys a particularly strong bargaining position in any negotiations: its leaders are well aware of the grassroots pressures on the Western leadership to negotiate reductions in strategic arms, while they themselves do not feel encumbered by such pressures....

Whether the Strategic Defense Initiative will ever be technically and economically feasible depends on the outcome of research and development now getting under way. The issue of whether it is politically desirable warrants much more consideration.

WILLIAM BAILEY

Oakton, Va.

To the Editors:

In the article "Space-based Ballistic-Missile Defense" the authors claim to have made a "highly optimistic" partial cost estimate of the power supply for a system designed to destroy 1,400 Russian ICBM's. Unstated is the assumption that all 1,400 missiles are launched in 100 seconds—that is, that they are all launched at the same time and proceed through the booster stage simultaneously. Even if this were a strategy implemented to thwart a satellite-based anti-ICBM system, it would be extremely difficult to accomplish. If one were to assume that the actual time available was on the order of 1,000 seconds (approximately 17 minutes), their "highly optimistic" cost estimate drops from $100 billion to $10 billion. The authors' estimate of the required number of "fighting mirrors" is similarly flawed.

Certainly the cost of a space-based anti-ICBM system would be high, but it distresses me to see a group of distinguished scientists make such an obviously impracticable assumption in an attempt to advance what is essentially a political argument.

DAVID R. CLARK

Garland, Tex.

To the Editors:

The arguments contained in the article "Space-based Ballistic-Missile Defense" appear to me to be fatally flawed on several counts. Some of my objections are as follows:

1. The authors specify that a defense must be perfect. Leaving aside the theological implications of perfection, I would argue that the question before us is whether an effective defense is possible. An effective defense is one that would make attack more difficult, more costly and more time-consuming, and that therefore makes the defender safer by dissuading an enemy from attacking. For example, any bank vault can be broken into with enough resources and time; vaults are built to dissuade safecrackers, not to defend deposits perfectly....

2. The reasoning underlying the conclusion that installation of ballistic-missile defenses would lead to more aggressive behavior strikes me as spurious. You could more reasonably argue that an adversary who was simultaneously facing a shield of uncertain effectiveness and protected by a shield of uncertain effectiveness would be doubly dissuaded from precipitous action: he would quite logically fear that his own defense would not work, whereas the other side's might.

3. The authors state that "if the president's dream is to be pursued, [outer] space will become a potential field of confrontation and battle." Frankly, I can't think of a better place to have a nuclear war than one where nobody lives....

4. The article advocates negotiating more treaties with the U.S.S.R. to bring about a safer world, without mentioning that the Soviets are violating the treaties that have already been negotiated. In the interest of honesty, the authors should have admitted that the strategy they advocate, of basing our security on pieces of paper rather than on strength of arms, is a very high-risk course that can succeed only if the Soviets depart from their established behavior.

EDWARD LAZEAR III

Louisville, Ky.

To the Editors:

We regret that space limitations and the inherent complexity of ballistic-missile defense (BMD) prevent us from addressing each of the points that have been raised in the many letters we have received.

To preclude the development feared by Mr. Bailey, in which the Soviets would have "a defense that could be very effective against [our] weak retaliatory blow," we have long supported programs that enhance the invulnerability of our strategic forces and provide our missiles with penetration aids. The status quo is one of rather stable deterrence, and such programs would maintain that equilibrium by enhancing our ability to retaliate after, and thus deter, a first strike.

As Bailey points out, the Soviets do have vast air defenses. Nevertheless, the Air Force is confident (as are we) that our manned bombers could penetrate to Soviet targets. This illustrates the prodigious task of BMD, for a bomber is a far slower, larger and more fragile target than a ballistic missile or a warhead.

Bailey worries that "the Soviets may not be very interested in equal treaties." We would say both sides have sought treaties that favor them-

Masterslice and the Engineering Design System

IBM

To: Abby
From: Roger
Subject: IBM Technology

I've been reviewing some of our past and present technological achievements, and it occurred to me that the scientific, engineering, and academic communities might like to know more about them. Will you select a topic from the following list?

Thanks.

- Vacuum tube digital multiplier
- IBM 603/604 calculators
- Selective Sequence Electronic Calculator (SSEC)
- Tape drive vacuum column
- Naval Ordnance Research Calculator (NORC)
- Input/output channel
- IBM 608 transistor calculator
- FORTRAN
- RAMAC and disks
- First automated transistor production
- Chain and train printers
- Input/Output Control System (IOCS)
- STRETCH computer
- "Selectric" typewriter
- SABRE airline reservation system
- Removable disk pack
- Virtual machine concept
- Hypertape
- System/360 compatible family
- Operating System/360
- Solid Logic Technology
- System/360 Model 67/Time-Sharing System
- One-transistor memory cell
- Cache memory
- Relational data base
- First all-monolithic main memory
- Thin-film recording head
- Floppy disk
- Tape group code recording
- Systems Network Architecture
- Federal cryptographic standard
- Laser/electrophotographic printer
- First 64K-bit chip mass production
- First E-beam direct-write chip production
- Thermal Conduction Module
- 288K-bit memory chip
- Robotic control language
- Masterslice and the Engineering Design System

Roger-
Let's tell about our innovative method of designing and integrating logic chips into our large computers.
Abby

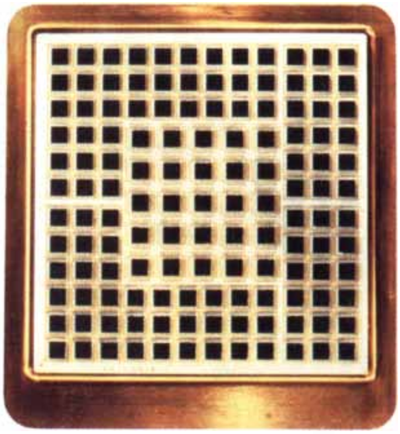
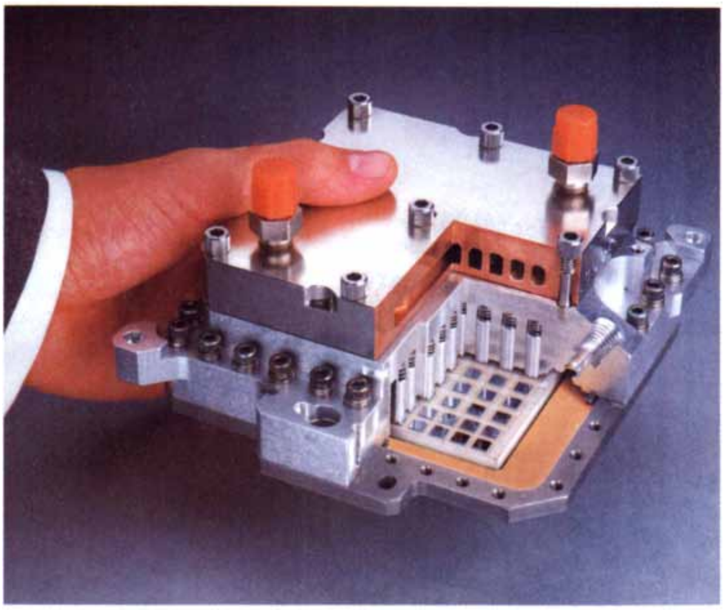


Figure 1: *The logic module used in large IBM computers (cutaway below) is part of the industry's densest circuit packaging. The electronic chips mounted in each module (right) were made through IBM's Engineering Design System and the masterslice concept: customize where necessary, standardize where possible.*



As computer applications continue to expand, designers of large computers are faced with many challenges. One of the biggest of these is designing semiconductor chips: not only do engineers have to design chips to contain the desired function, but they must also integrate the chips into the rest of the system and accomplish this quickly and inexpensively.

For nearly two decades, IBM designers have been leaders in this field, pioneering the technologies of chip customization, automated design, and automated manufacturing. In the mid-1960s, IBM researchers began developing a chip customization technology—known as gate array or masterslice—as well as a totally integrated set of design automation tools called the Engineering Design System.

The first masterslice chip came off IBM production lines in 1967 and was part of the System/3 announced in 1969. Growing increasingly important as an element in IBM computers, masterslice became the basis for the logic in the System/38 in 1978. This marked the first major impact of masterslice technology on computer architecture, making masterslice a driving force in semicustom, large-scale integration of chips in the computer industry.

In masterslice, a predefined pattern of circuit elements is fabricated in an area of a silicon chip called a cell. The pattern is then repeated so that almost the entire chip is covered with identical cells. In this manner, many chips

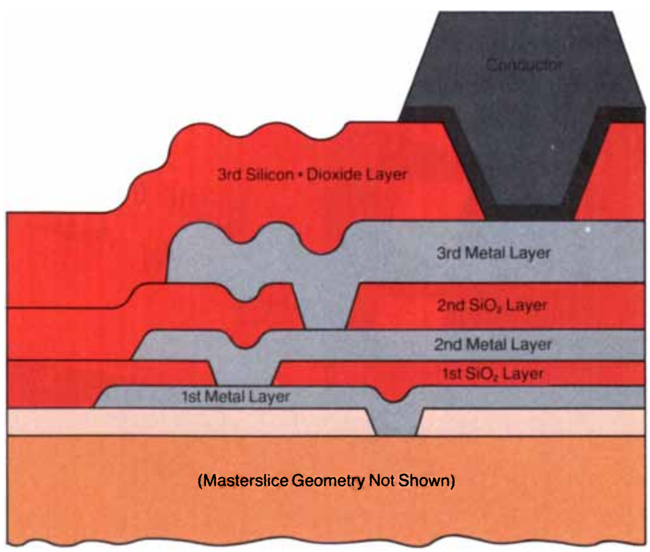


Figure 2: *This simplified side view of a logic chip shows three layers of metallization (along with three layers of insulating silicon dioxide) that are put on top of the masterslice to produce a semicustom chip. The metallization process enables designers to customize chips for a specific job. And a standard "base"—the masterslice—allows quicker turnaround times and lower manufacturing costs.*

Masterslice and the Engineering Design System

may be produced with identical arrays of identical cells.

Customization takes place in "metallization"—the adding of alternate layers of insulators and metal wiring interconnections over the masterslice pattern of the circuit elements. This gives chip designers the freedom to make hundreds of variations in their design and still maintain the economic standardization of parts.

Masterslice technology has grown into an important process for implementing logic in IBM products. It is the basis of the 1,200 logic chips that make up the 500 different logic configurations of the central processing unit of IBM's largest computers, the 308X family.

IBM's Engineering Design System (now a full family of integrated design tools) has a data base that contains a complete description of each chip and its relation to the rest of the system, from the physical properties of individual devices to the requirements of the entire logic system of the computer. Thus, this design system enables

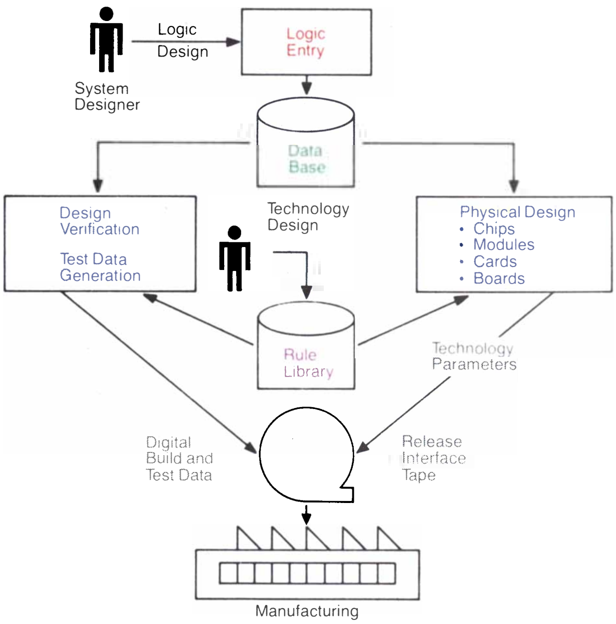


Figure 3: *With IBM's Engineering Design System, machine designers use terminals to input logic functions for a chip and establish a data base. Through simulation, the system provides logic verification and performs logic delay checking. Test patterns are then automatically generated for each part. In the meantime, physical design of the chip is done with computer programs that perform the following tasks: circuit placement, I/O assignment, wiring, and checking. All physical design information is then transformed into shapes, patterns, and precise locations of interconnections and circuit elements required for manufacturing.*

Figure 4: *Shown here is a display screen from the Interactive Graphic System (IGS), one of the many Engineering Design System tools developed by IBM to speed chip development and implementation. IGS is a powerful shape manipulation tool used to design new masterslices.*

the needs of a large system to be reflected in the design of its smallest components.

The thousands of individual software modules of the Engineering Design System can be used to take a chip from initial design, through simulation and testing, to manufacture. Linking such a wide range of functions through common interfaces to form a total system is a feat unmatched in the industry. A designer using this system can take a chip from the start of the physical design stage to the manufacturing line in about six days.

Many engineers, scientists, and programmers throughout IBM contributed to the development of masterslice and the Engineering Design System. Their contributions are only part of IBM's continuing commitment to research, development, and engineering. IBM®

selves, a quite natural trait that has hindered and stymied progress. Nevertheless, the two governments have sometimes hit on valuable agreements that they deemed more desirable than no agreement at all. It is now the announced intent of President Reagan to seek such an agreement.

In answer to Mr. Clark, our explicit assumption that the Soviets could launch all their current 1,400 ICBM's at once is designed not "to advance... a political argument" but only to acknowledge a Soviet capability. The Strategic Defense Initiative (SDI) program makes the same assumption. Indeed, our defense is likely to face a much larger force than we assumed in our article, for the Soviets have repeatedly stated that they view SDI as an attempt by the U.S. to gain strategic superiority and that they will not accept such an outcome.

In that connection we address a crucial question raised by Mr. Lazear: "An adversary who was simultaneously facing [our] shield of uncertain effectiveness and protected by [such] a shield would be doubly dissuaded...." True enough, provided he cannot enlarge or improve his offensive force. But an uncertain defense would destroy all negotiated limits on the offense: by the same logic, he will feel doubly threatened by us, and will therefore feel compelled to enlarge his offensive force in order to maintain a deterrent in which he has confidence. And then we, being keenly aware of the ineffectiveness of our own defense, will feel threatened by his buildup and react accordingly. (The uncertainty inherent in strategic defense would also make good-faith arms control vastly more difficult than it has been for offensive forces, whose capabilities are well known and agreed on.)

Mr. Lazear suggests we ask "whether an effective defense is possible." But time does not stand still while the U.S. deploys strategic defenses, and although the U.S. is wealthier than the U.S.S.R., it does not have vastly more resources to spend on strategic defense than the Soviet Union has available to overcome such defense. Therefore supporters and critics of strategic defense alike consider the "cost-exchange ratio," which is unity when the defense can be overcome with the expenditure of equal resources. Defense of Minuteman silos can make attack relatively more costly, but defenses of cities can be overcome more readily than they can be deployed. Unless the cost-exchange ratio is favorable, the situation is likely to be worsened by deployment of strategic defenses. Such supporters of strategic defense as Edward Teller also insist on a favorable cost-exchange ratio before deployment of a defense.

Mr. Lazear asserts that "in the interest of honesty" we should admit that Soviet behavior makes arms control "a very high-risk course." In assessing any arms-control proposal one must ask whether our security would be under greater threat from clandestine treaty violations that escape detection or from the activities that would legitimately occur in the absence of a treaty. By this criterion all the SALT accords have enhanced our security, as would the space-arms-control regime we advocate. We say this even though we believe the Soviet radar under construction at Krasnoyarsk will be the first serious violation of a ratified strategic-arms-control treaty if it is completed. This radar would be of little military significance, but it is politically essential that the U.S.S.R. abandon it or modify it to bring it into treaty compliance.

We do not agree with Lazear that uninhabited outer space is the ideal place for nuclear war, for such a war would quickly spread to the vital ground-control facilities. In all likelihood it would then escalate to general nuclear war.

We stress that one should not lose sight of the crux of this issue by focusing on the technological details of the BMD schemes under discussion, since none of them hold out the hope of providing a sturdy shield against a determined nuclear attack on the nation. That fact has been recognized by many advocates of SDI, who now base their case on the contention that strategic defense will bolster deterrence. But that abandons President Reagan's original vision—the hope that SDI would replace assured destruction by assured survival. A defense that is merely an adjunct to a preponderant offensive deterrent brings us back to the very point where the superpowers were in 1972 when they recognized that BMD and offensive-arms limitations are incompatible and acted on that insight by signing the ABM Treaty. We therefore welcome President Reagan's decision to initiate negotiations on space weapons *and* offensive missiles. An accord that encompasses both would be in our own security interest, but it will not be attainable if we refuse to face the implications of pursuing strategic defense.

HANS A. BETHE

RICHARD L. GARWIN

KURT GOTTFRIED

HENRY W. KENDALL

A year's worth of reports, plans, schedules, charts, graphs, files, facts and figures and it could all be lost in the blink of an eye.

The most important part of your computer may be the part you've considered least—the floppy disk. After all, there doesn't seem to be much difference between one disk and another. But now Fuji introduces a floppy disk that's worth a second look.

We designed our disk with the understanding that one microscopic imperfection can erase pages of crucial data. That's why every Fuji Film Floppy Disk is rigidly inspected after each production process. And that's why each one is backed with a lifetime warranty.

We've even considered how carefully a disk has to be handled, so we designed user-friendly packaging that makes it easier to get the disk out of the box. And we provided plenty of labeling space, so you won't have any trouble telling which disk is which.

So think twice before buying a floppy disk. And then buy the one you won't have any second thoughts about.

Fuji Film Floppy Disks.

FUJI.

Nobody gives you better performance.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

FEBRUARY, 1935: "Savings of millions of dollars are accruing annually to industry through the use of cemented carbides for dies, tools and wear-resistant parts. These metallic substances of near-diamond hardness have virtually revolutionized the extruding, drawing, cutting and machining of metals in mass production by making possible an extraordinary stepping up of operating speeds with marked improvement in quality. Cemented carbides are carbide particles cemented together, cobalt and sometimes nickel being used as the binder. The carbides are those of tungsten, tantalum, titanium and molybdenum."

"The archaeological survey carried on by the Tennessee Valley Authority in the Norris and Wheeler dam reservoir areas during the past year constituted an attempt to bring to light before those areas were submerged any data on prehistoric records. Some 40 Indian mounds were uncovered. Each mound marked the site of a village, and in the center of each a town house or temple, ranging anywhere from 40 to 60 feet long by 30 to 35 feet wide, was in evidence. The site of an early fortified village was unearthed near Caryville."

"Science today touches and influences practically every material object. It would not seem, however, that architecture would show that influence strongly. Building design and construction was essentially the same a very few years ago as it was when the ancient inhabitants of Ur discovered the principle of the arch and the dome. Today steel frames and the elevator have started the trend toward skyscrapers. Much experiment has been carried on in the development of monolithic concrete houses. Steel houses and houses employing glass building blocks have also been widely discussed. Perhaps one of the greatest influences now pleading for close study as to its relation to building design is air conditioning."

"A long-range radio compass that enables itinerant pilots flying over oceans and unknown land to find their way directly to any radio broadcasting station has been developed by the Westport Manufacturing Company. The radio compass consists of an indicating dial, a compass converter unit, a dynamotor and a loop antenna. A pilot may tune in on a broadcast station. If the plane is headed directly toward it, the indicator needle points to zero. If the plane is off course ever so slightly, the needle points right or left as the case may be."

SCIENTIFIC AMERICAN

FEBRUARY, 1885: "There is an excitement in the celestial court. Encke's comet has arrived, and star gazers are turning their telescopes to the skies in eager haste to obtain a glimpse of the distinguished visitor. Our eccentric guest is not a prince among comets. It does not burst upon the astonished gaze at noonday with a brilliancy akin to that of the sun; its tail is not curved like a Turkish cimeter, nor does it branch out into six tails, each one 6,000,000 miles long. Encke's comet is interesting chiefly for being the first known comet of a short period, for making the shortest circuit of any member of its class, for performing its revolution within the boundaries of the solar system, and for the reason that it seems to be more amenable to physical law than some of the more imposing members of the cometary family."

"Has electric light any effect on the growth of plants? Isaac Buchanan, one of the best-known florists of New York, stated that his observation inclined him to believe that when plants were used for decorative purposes in rooms where the electric light was used instead of gas, they seemed to have all the health and vigor as if growing under the light of a conservatory. Few commercial florists have the means or time for such expensive experiments as would be necessary to determine whether the use of the electric light in forcing flowers and fruit in greenhouses during winter could be profitably employed. It is a matter of sufficient importance, it would appear, for the Agricultural Department at Washington to take hold of."

"The U.S. Hydrographic Office, in pursuance of its policy to lessen the dangers of navigation, has recently commenced the collection of information to determine the best manner of using oil to calm the surface of troubled waters. This matter has long been the subject of controversy. In 1844 a Dutch commission, after pouring a few gallons of oil on the storm-beaten bosom of the North Sea, and finding the waves not sensibly affected, declared that the oft-repeated account of the saving of ships by this means was a fantastic creation of the imagination. Notwithstanding this, Scotch coasters have saved themselves again and again by strewing the sea with the fatty parts of fish, cut into small pieces, which were carried with them for the purpose; and so much reliable information on this subject has now been collected from the common experience of seafaring men, that the evidence in its favor can no longer be controverted. It must be understood, however, that the application of oil does not make the surface perfectly smooth, but merely lessens the effect of what the seaman calls 'combers,' or the great, broken, rolling masses of water which have first disabled and then swamped so many ships."

"Eugene Boban, an antiquarian and a traveler, well known to anthropologists and ethnographers, has made a restoration of the costume of a young Aztec warrior, a chevalier of the army of Montezuma. This truly remarkable object is designed for one of the largest ethnographic collections of Mexico. The Mexican warrior of the epoch of the conquest (1521) was, as may be seen, clad in a material spotted to resemble the skin of a tiger [*sic*], and wore a helmet shaped like the head of that animal."

Eugene Boban's Aztec reconstruction


GM
Visit the Riviera this year.
It's the trip of a lifetime. With its timeless styling and seemingly endless list of standard appointments, the Riviera is a classic.
Yet the Riviera is as contemporary as today. With front-wheel drive and independent four-wheel suspension. On the T Type, there's a 3.8-litre turbo-charged V-6 with sequential-port fuel injection (SFI).
And you can make the Riviera even more unordinary. Just choose the available trim package, with leather and suede seating areas, genuine burled walnut veneer instrument panel and wood/leather steering wheel.
Buckle up and visit your Buick dealer.
To ask any questions, request a brochure or test drive, call the Buick Product Information Center, 8 a.m. to 8 p.m. Eastern time, weekdays:
1-800-85-BUICK (1-800-852-8425).
Wouldn't you really rather have a Buick?
BUICK
The Buick Riviera: a contemporary classic.

THE AUTHORS

ROBERT L. WESSON and ROBERT E. WALLACE ("Predicting the Next Great Earthquake in California") work in the U.S. Geological Survey's Office of Earthquakes, Volcanoes and Engineering, Wesson at its Reston, Va., office as a geophysicist and Wallace in Menlo Park, Calif., as its chief scientist. Wesson did his undergraduate work at the Massachusetts Institute of Technology and earned his M.S. and Ph.D. degrees at Stanford University in 1968 and 1970 respectively. In 1970 he joined the Geological Survey as a research associate. His assignments since then have included earthquake-prediction field studies along the San Andreas fault and in Soviet Central Asia, and he has held the posts of chief of the Office of Earthquake Studies (1978 to 1980) and assistant director of the Geological Survey (1980 through 1982). In 1982 Wesson returned to full-time research. Wallace is a graduate of Northwestern University and the California Institute of Technology, where he got a Ph.D. in 1946. While doing his doctoral work he held a position with the Geological Survey; in 1951 he joined the Survey on a permanent basis. From 1972 to 1977 Wallace was chairman of the U.S. working group of the U.S.–U.S.S.R. Cooperative Program on Earthquake Prediction.

JOHN E. DONELSON and MERVYN J. TURNER ("How the Trypanosome Changes Its Coat") are respectively professor of biochemistry at the University of Iowa and a fellow at the Medical Research Council's Molteno Institute at the University of Cambridge. Donelson holds a B.S. from Iowa State University and a Ph.D. in biochemistry from Cornell University. In 1974, after postdoctoral work at Cambridge and at Stanford University, he joined the faculty at the University of Iowa. His interest in tropical parasitic diseases originated during a stint in Ghana as a Peace Corps volunteer between college and graduate school; his involvement took him abroad again in 1980, when he traveled to Kenya as a visiting scientist at the International Laboratory for Research on Animal Diseases. Donelson is currently a Burroughs-Wellcome scholar in molecular parasitology. Turner is a graduate of the University of Sheffield, where he got a Ph.D. in organic chemistry in 1970. He pursued his interest in the biochemistry of membrane proteins as a postdoctoral fellow at Harvard University (1971 to 1974) and at Queen Victoria Hospital in London (1974 to 1977). In 1977 he took his present position at the Molteno Institute. In April, Turner will assume the directorship of a new biochemical-parasitology program at the Merck Sharp & Dohme Research Laboratories in Rahway, N.J.

BRADLEY E. SCHAEFER ("Gamma-Ray Bursters") is a research associate in the Laboratory for High-Energy Astrophysics at the Goddard Space Flight Center in Greenbelt, Md. With a one-year interruption for graduate work at the University of California at Berkeley he received both his undergraduate and his graduate training at the Massachusetts Institute of Technology, which granted him a Ph.D. in physics in 1983. Other interests include cataclysmic variable stars and the history of ancient Chinese astronomy. When he is at leisure, Schaefer plays chess, in which he recently attained the rating of expert.

BERNARD H. LAVENDA ("Brownian Motion") is professor of chemical physics at the University of Camerino. He studied chemistry at Clark University in Massachusetts, the Weizmann Institute of Science in Israel and the Free University of Brussels, where he got his doctorate in 1970. After a two-year postdoctoral fellowship at the Hebrew University of Jerusalem he taught at the universities of Pisa and Naples. Lavenda joined the faculty at Camerino in 1980.

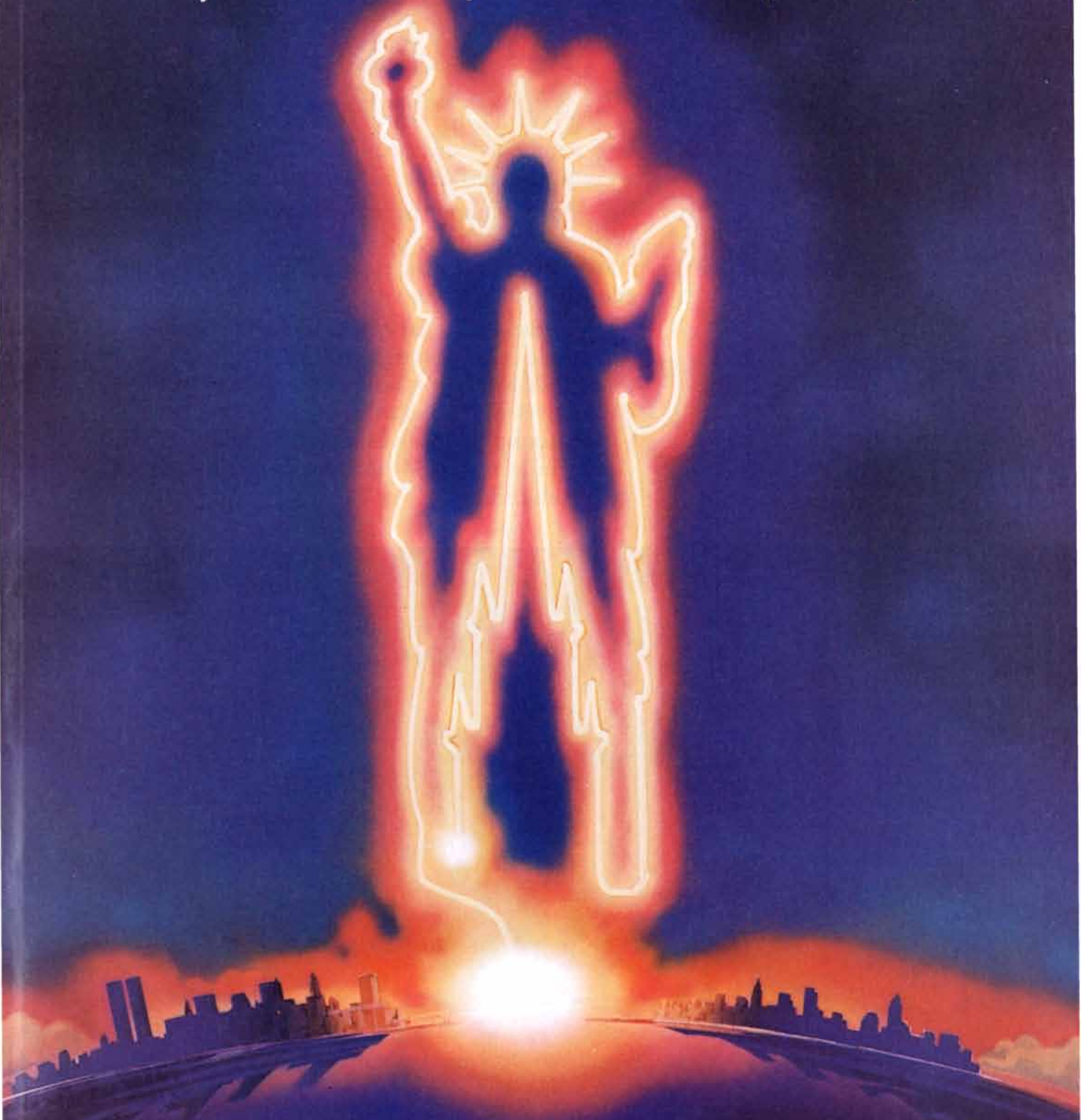
T. H. CLUTTON-BROCK ("Reproductive Success in Red Deer") is a Royal Society Research Fellow in the department of zoology at the University of Cambridge. He did undergraduate and graduate work at Cambridge, getting his doctorate in 1972 for a study of the ecology and breeding system of red *Colobus* monkeys. In 1976, after a year at the University of Oxford as a research fellow and three years at the University of Sussex as a lecturer in ethology, he returned to Cambridge to join the King's College Sociobiology Group. In 1980 Clutton-Brock established the Large Animal Research Group within Cambridge's zoology department.

ELI BROOKNER ("Phased-Array Radars") is a consulting scientist at the Raytheon Company's Equipment Division in Wayland, Mass. He completed his undergraduate work in electrical engineering at the City College of New York in 1953 and joined the Columbia University Electronics Research Laboratories, where he took part in studies of microwave noise and in the design of long-range radars. He remained there until 1962, with an interruption from 1957 to 1960 for work as a project engineer at the Federal Scientific Corporation in New York. Simultaneously he pursued graduate studies at Columbia, earning his M.S. in 1955 and his D.Sc. in 1962. In that year Brookner joined Raytheon, where he has helped to develop numerous defense radars and has participated in studies of space-based radar systems.

GERALD E. LOEB ("The Functional Replacement of the Ear") is permanent senior investigator in the Laboratory of Neural Control at the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) in Bethesda. He holds a B.A. in human biology (1969) from Johns Hopkins University; his M.D. (1972) is from the Johns Hopkins University School of Medicine. While in medical school he was a research assistant in Johns Hopkins' department of biophysics, the principal investigator in a biomaterials study and a guest research associate at the University of Utah's artificial-eye project. From 1972 to 1973 he was an intern in the department of surgery at the University of Arizona College of Medicine; in 1973 he joined the NINCDS. Loeb is president of Biomed Concepts, Inc.

OLEG D. SHERBY and JEFFREY WADSWORTH ("Damascus Steels") are respectively professor of materials science and engineering at Stanford University and a staff scientist at the Lockheed Palo Alto Research Laboratory. Sherby is a graduate of the University of California at Berkeley, which granted him a Ph.D. in 1956. He spent seven years as a research metallurgist at the Institute of Engineering Research at Berkeley, then went to England for two years, working first as a postdoctoral fellow at the University of Sheffield and then as a scientific liaison officer with the U.S. Office of Naval Research. He joined the Stanford faculty in 1958. Sherby's research interests lie in mechanical behavior, diffusion phenomena and the processing of materials. Wadsworth holds a Ph.D. from Sheffield. In 1976 he took the post of research associate at Stanford, and he continues his association with the university as a consulting associate professor. At Lockheed, where he has worked since 1980, he is developing superplastic aluminum alloys and heat-resistant metals for use in aircraft and aerospace structures.

Without us, the home of free speech could lose the power of speech.



In the communications industry, a leading buzzword for the future is fiber optics. At NYNEX, it's a word for yesterday. Because NYNEX has 50,000 miles of these hair-thin fibers in place now.

In the most information-hungry communities in the country: Boston, New York, and their surrounding areas.

It's one of the world's largest lightwave installations. It transmits calls, data, and saves space for the future in congested tunnels.

Water can't harm it. And neither can extreme temperatures.

In short, it's a network tough enough to make sure the home of free speech never loses the power of speech. It's what you get when you team New England determination with New York foresight.

NYNEX is the parent company of New York Telephone and New England Telephone, plus subsidiaries that offer mobile services, directory publishing and business communication equipment.

For information, call (800) 441-1919. In Utah, call (800) 453-9000. Or write T. Parra, Dir. Investor Relations, NYNEX, Box 2945, NY, NY 10185.

Tough demands breed tough minds.

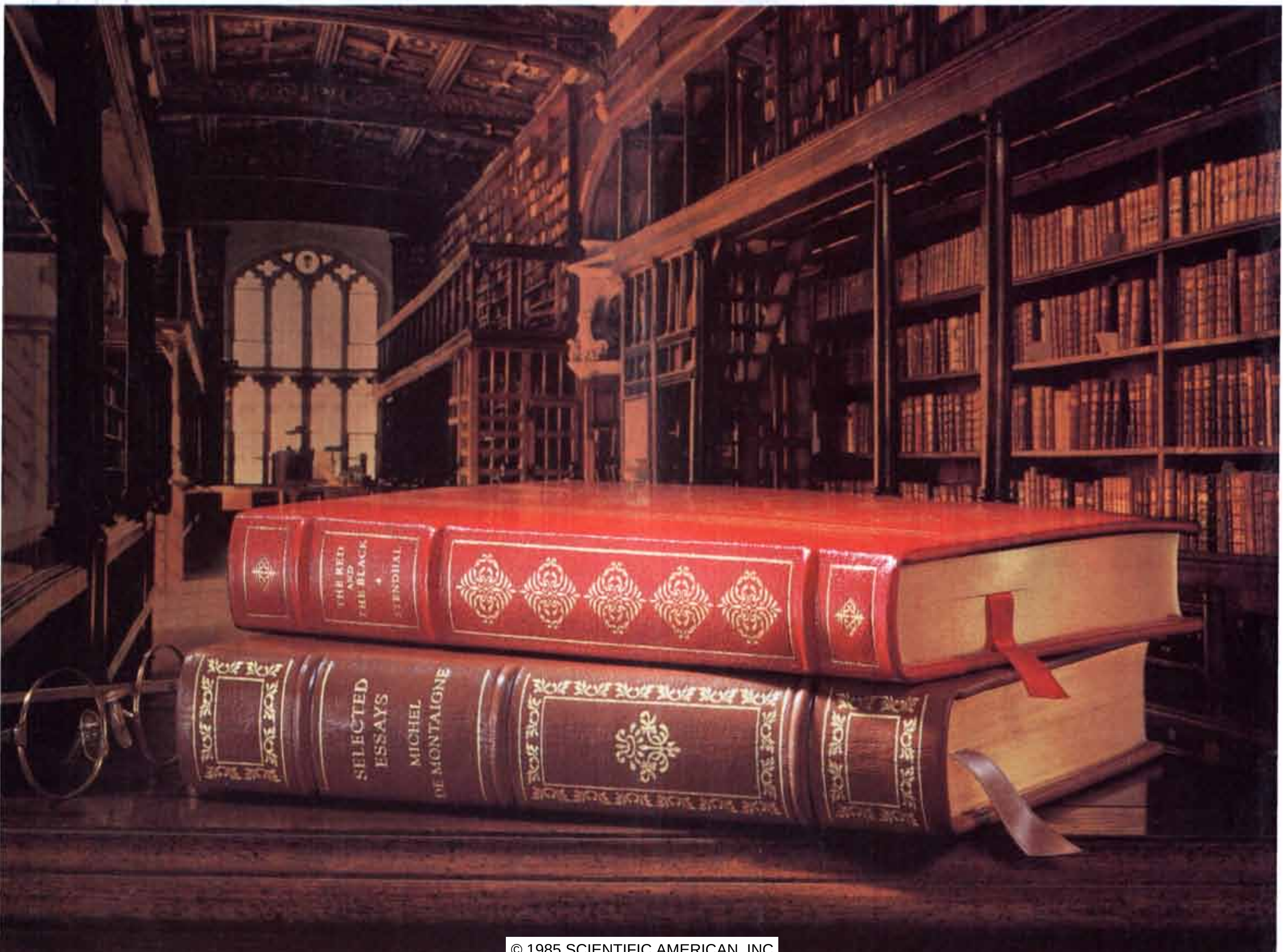
NYNEX

The Oxford Project

Oxford University. Here, in this ancient seat of learning, a board of eminent Oxford dons recently entered upon a truly historic task.

From among the many thousands of books written through the ages, they selected the 50 most important—the supreme achievements in the world of literature.

These works have been brought together in a single collection to form 'The Oxford Library.' And this collection is now being offered in a most distinctive edition, so that people everywhere can surround themselves with the finest writing—the finest reading— that man's literary genius has produced.



COMPUTER RECREATIONS

An expert system outperforms mere mortals as it conquers the feared Dungeons of Doom

by A. K. Dewdney

Every year thousands of people die a fantasy death in the Dungeons of Doom while playing Rogue, one of a new generation of computer adventure games. The player watches a map of the dungeons on the display screen and directs the actions of his or her game persona, a character called the Rogue. The object of the game is to descend through the 26 levels of the dungeons, seize the Amulet of Yendor and return safely to the surface, gathering gold and slaying or escaping from monsters along the way. Few human players survive the dangers of this subterranean odyssey, let alone return with the amulet.

The fantasy can be quite vivid; one quickly leaves awareness of keyboard and display screen far behind. As the Rogue, I approached the entrance to the dungeons with some trepidation: it was night and the ancient ruins that marked the point of my imminent descent had a gloomy and forbidding appearance. In preparation I set out my enchanted mace, a bow and a quiver of arrows snatched from a dragon's hoard in the distant Dark Mountains. I donned my elf-crafted armor, picked up my weapons and food and stepped into the stygian darkness of a stairway.

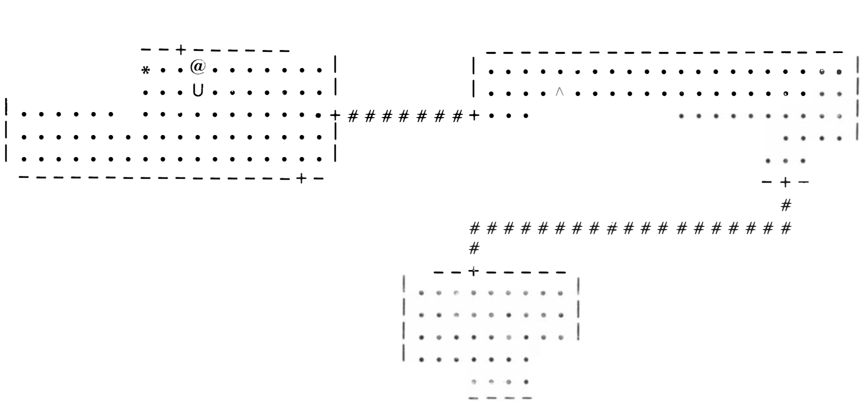
Just as the descent began to seem endless, I bumped into an oaken door and swung it cautiously open. There before me was the first room on the uppermost level of the Dungeons of Doom. It was dimly lighted by tapers and I trod carefully to the middle of the chamber, the better to survey it. Suddenly the floor under me gave way and for a breathless second I was falling. With a sickening thud I landed in another room. This one was much darker and, even when my eyes became accustomed to the dark, I could see only a few feet in any direction. To make matters worse, I could hear something moving about in this chamber. Nauseous with fear, I stumbled to my feet to find myself confronting a squat, horrid armor-clad figure holding a club. As it raised its weapon to strike, a rush of adrenalin cleared my head. I fitted an arrow to my bow, drew it and fired all in one fluid motion. (Fortunately I took archery as a noncredit course in my senior year.) There was a swish, a thwock and a squeal as the goblin (for that is what the creature was) fell to the floor quivering with unvented rage. I stepped gingerly away from it, determined to find a stairway up and out of the Dungeons of Doom. I thought of my cozy home and my writing desk with the unfinished "Computer Recreations" article on it.

As I shuffled slowly in the direction in which I guessed a stairway to lie, my boot struck what felt like a small heap of stones. I looked down. Even in the dark something gleamed. Gold! It occurred to me to explore just one more room before returning to the surface....

Aside from the intensity of vicarious experience, Rogue goes a step beyond traditional adventure games in at least two respects. First, the layout of the terrain is generated by the program itself and the layout changes from one game to the next. Second, Rogue supplies the human player with a plan view of the current level of dungeons as explored so far [*see illustration on this page*]. Various features are indicated by different keyboard characters drawn on the screen. For example, bars and hyphens represent walls, number and percent signs represent passages and stairways respectively, plus signs represent entrances and carets represent traps. Objects and features of the chamber may or may not be visible depending on whether or not the chamber is illuminated. When the Rogue enters this particular level, no features are shown on the screen. In order for them to be represented the Rogue (represented by an "at" sign) must discover them. And so he explores, stumbles on some traps, scouts parts of various rooms and traverses passageways. He may encounter the letter *U,* which represents a monster known as the Umber Hulk.

Below the character-oriented graphic plan the screen displays the Rogue's current statistics. In the illustration the Rogue is on the 25th level, he has accumulated 7,730 pieces of gold, he has earned 77 hit points (but only 25 remain because of damage sustained), he has accumulated strength to the 15th degree (18 is the highest), he wears class-9 armor and has 30,668 experience points, enough to place him on the exalted 13th order of experience. At the 25th level of the Dungeons of Doom this Rogue has only one more stairway to descend before reaching the bottom level and attempting to snatch the coveted Amulet of Yendor. The Rogue must first dispose of the Umber Hulk, however.

This discussion explores only the surface of the game. It would be possible to write an entire book of rules and advice for playing Rogue. So far the only available document of that nature is a terse, eight-page report written by Michael C. Toy and Kenneth C. R. C. Arnold, *A Guide to the Dungeons of Doom* [see "Bibliography," page 128]. Unfortunately it is available only to users of the VAX/UNIX time-sharing system. Arnold tells me that a version of Rogue suitable for the IBM PC is now available from A.I.



The 25th level of the Dungeons of Doom

Design, 201 San Antonio Circle, Suite 115, Mountain View, Calif. 94040. Toy and Arnold, Rogue's creators, evidently think the best way to learn the game is to play it. Nevertheless, the general features of the game can be delineated.

At each level the playing area is divided into squares. The Rogue occupies one square at a time as he explores his vicinity. Movement within rooms and passages is controlled by typed letter commands such as *h, j, k* and *l* that move the Rogue a single square in one of the four main directions. Other commands produce diagonal movement or continued motion in a given direction. To make the Rogue ascend or descend a stairway, the player must type the character < or >.

When an object is discovered near the Rogue's current location, moving the Rogue to the square containing the object automatically causes the object to be picked up. If the player wants to move onto that square without picking up the object, he or she must type *m* followed by the appropriate direction character. It is also possible to have the Rogue search for traps by typing *s*, an action that effects a search of neighboring squares. A caution is in order: this search has only a 20 percent chance of discovering a trap.

Occasionally the Rogue must rest (type a period) or eat (type an *e*) to regenerate strength expended in wandering about or in subduing monsters. Apart from a limited amount of food carried in a backpack, the Rogue has nothing else to eat except what can be found on the dungeon floors (unappetizing as this may sound).

When a piece of armor is found, the Rogue may pick it up (automatically placing it in the backpack) or wear it (type a *W*). Armor naturally gives the Rogue more protection in combat, but armor may be cursed. If the armor is cursed, efforts to take it off (type a *T*) are useless without a magic scroll that breaks the curse. Any magic rings found by the Rogue may be put on (type a *P*) or removed (type an *R*) unless they are cursed.

At times it becomes necessary for the Rogue to drop a flask and a scroll or two (type a *d* followed by an object character) because his pack is full. (A Rogue in full pack cannot pick up armor.) Before dropping a flask, however, the Rogue might benefit from drinking the unknown contents: some potions have a healing effect and others enable the Rogue to see the Invisible Monster. On the other hand, the potion may merely cause confusion, so that if one types a command for the Rogue to move north, he may wander off in a random direction. By the same token, there may be an advantage to reading a scroll (type an *r*) before discarding it: certain formulas remove the curse from one's armor.

@	the Rogue	h	move one space west
–|	walls of rooms	j	move one space south
+	doorway	k	move one space north
#	passage between rooms	l	move one space east
·	floor of room	>	descend stairway
%	staircase	<	ascend stairway
^	trap	m	move onto something
)	weapon	s	search for trap
]	piece of armor	^	identify trap
*	gold	. /e	rest/eat
!	flask of potion	W/T	wear/take off armor
?	magic scroll	P/R	put on/remove ring
:	food	d	drop an object
/	magic wand	q	quaff a potion
=	magic ring	r	read a scroll
A–Z	uppercase letters denote denizens of dungeons	z	zap with magic wand
		t	throw an object
		w	wield a weapon
		f	fight to the death

Map symbols **(left)** *and commands for the game of Rogue*

The most impressive magic devices at the Rogue's disposal are wands. Depending on the type of wand picked up, the Rogue may zap a monster (type a *z*) with various effects: he may transport it to a random location, shoot fireballs at it or change it into another monster. The last type of wand, called a polymorph wand, is best used on the more horrible monsters: it is much better to transform the dread Purple Worm into a bat than vice versa.

When the Rogue discovers a monster, a good policy might be to leave it alone. Sometimes a monster is sleeping and will not attack if left undisturbed. If a fight seems inevitable, however, one may wield a weapon (type a *w*) and then fight to the death (type an *f*). This can be done only after one has moved the Rogue next to the monster. The outcome of the struggle will depend probabilistically on such factors as the Rogue's current strength level, degree of experience and armor class.

Foremost in the ranks of players of Rogue stands a top-caliber nonhuman player. For the past four years a computer program has been playing the game and matching the prowess of the best Rogue runners in the quest for the amulet and gold; it provides as well an intriguing opportunity to watch an expert system at work.

On February 16, 1984, at the University of Texas at Austin, a Rogue manipulated by a program fought off all monsters, amassed a considerable pile of gold and returned with the amulet. Bearing the distinctly un-Tolkienesque name of ROG-O-MATIC, this program directed the Rogue's every step, every pause, every throw and blow.

ROG-O-MATIC is the creation of four graduate students in the Computer Science Department at Carnegie-Mellon University in Pittsburgh: Andrew Appel, Leonard Hamey, Guy Jacobson and Michael Mauldin. ROG-O-MATIC links programmed knowledge sources with expert systems to make decisions about what to do in every conceivable underground situation.

When a human plays Rogue, a stream of commands from the keyboard flows by way of the operating system into the ROGUE program. The program automatically decides which monster to confront the Rogue with next, which layout to use for the next level of dungeons, and so on. The program transmits this information, again by way of the operating system, to the display screen in order to keep the human informed of the current situation.

In replacing a human player, the ROG-O-MATIC program intercepts the keyboard character stream and transmits characters of its own to the ROGUE program. The latter has no idea (so to speak) that another program is playing. By the same token, information directed by the ROGUE program to the screen is also directed toward the ROG-O-MATIC program so that it can keep its own map of the Dungeons of Doom.

The ROG-O-MATIC program consists of 12,000 lines of the programming language C; it is even longer and more complicated than the ROGUE program. ROG-O-MATIC began in 1981 as what its original creators, Appel and Jacobson, thought would be a "simple project." Shortly after Mauldin joined the group the early Rogue-playing program went through an extensive round of modifications, each version adding some enhancement in tactical or strategic play. By the time the fourth member, Hamey, was contributing to program development, it was beginning to dawn on the authors that they had created, in effect, an expert system. Such systems constitute a key application of the so-called Fifth Generation of computers, projected to arrive in the marketplace in the late 1980's. An expert system is designed to embody and project human expertise in a variety of areas from medicine to engineering.

By using the type of software architecture that is suited to expert systems, the creators of ROG-O-MATIC were able to design and modify their program with relative ease. In particular, they organized the kinds of knowledge and expertise required by the Rogue into a hierarchy of various subsystems [*see illustration below*].

For example, one high-level expert (called Melee) controls fighting during combat and another high-level expert (called Target) directs the Rogue's pursuit of monsters. Both these experts use a lower-level expert called *battle,* which carries out special attacks or initiates retreat as the situation demands. The battle expert thus sometimes calls on the retreat expert for help, and the latter invariably draws on a source of knowledge called *pathc.* This is a special algorithm that searches the local terrain for the shortest path to whichever location is specified. Shortest-path algorithms are well developed in general, and in this particular application only the fastest algorithm possible is acceptable; it is used almost continuously as ROG-O-MATIC explores the Dungeons of Doom. Knowledge about terrain used by *pathc* comes from *termap,* essentially a data structure recording the terrain features so far discovered by the Rogue as he explores a particular level of the dungeons. Finally, *termap* obtains all its knowledge from *sense,* a low-level data structure containing all the relevant output of the ROGUE program.

Before listing the duties of other experts and knowledge sources, I should like to reexamine one of the experts, *battle,* in more detail. Once a battle is under way, the melee expert calls on the battle expert to decide whether to attack or to retreat. Discretion being the better part of valor, the battle expert first determines the desirability and feasibility of retreat. In order to do so it must check some preconditions:

1. The Rogue is not currently under the influence of the potion of confusion (which could cause the Rogue to flee directly toward the monster).

2. A monster is not already holding the Rogue fast.

3. It would be possible to die in one melee round (avoiding conflict would thus be highly desirable).

4. The retreat expert can find an escape route (a retreat is possible).

If all four preconditions are met, ROG-O-MATIC will turn the matter of the Rogue's retreat over to the retreat expert. If they are not met, the battle expert now runs through a list of aggressive possibilities.

1. If it is possible for the Rogue to die in one melee round, if the monster is nearby and visible to the Rogue and if the Rogue happens to have a teleportation wand, then point the wand at the monster.

2. If it is possible for the Rogue to die in one melee round, if the monster is nearby and if the Rogue has a teleportation scroll, then read the scroll's magic formula.

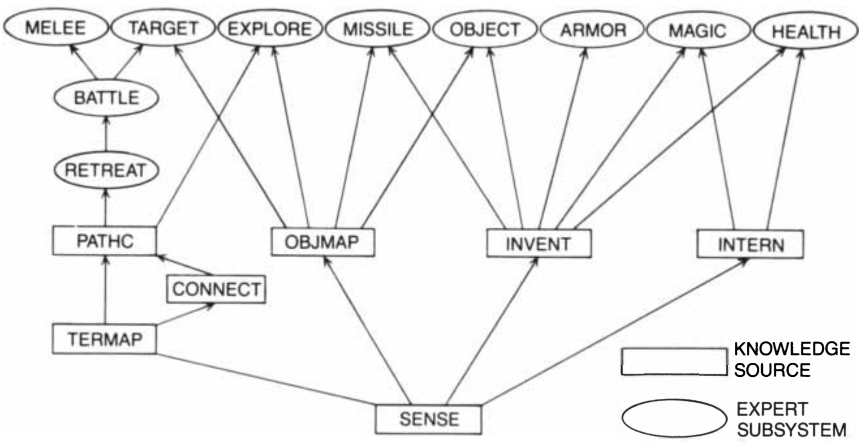
If neither of these conditions prevails in an encounter, there is no longer an alternative to going at least one round with the monster, and so the ROG-O-MATIC program cheerfully commits its Rogue persona to the battle. Perhaps the Rogue is strong enough to endure at least one melee round. Otherwise he may well be killed.

Other experts used by the program include an exploration expert (*explore*), which decides where to explore next and what motions to use. Also included are a missile expert, which handles the firing of arrows, rocks, spears and so on at monsters, an expert for choosing which objects to pick up, called *object,* an expert for choosing what armor to wear, an expert at using magic and an expert called *health,* which decides when to eat or rest. *Objmap* is a data structure that keeps track of the location and history of all objects so far encountered, *invent* is an inventory of items in the Rogue's backpack and *intern* is an internal-state recognizer that watches over the Rogue's readiness for exertion.

It is no surprise that ROG-O-MATIC can play a game of Rogue quickly. After a few minutes of computer time it is all over one way or the other. To date the program has played more than 12,000 games of Rogue at Carnegie-Mellon alone; statistics cited by ROG-O-MATIC's four creators tend to support the claim that the program's abilities now exceed those of the vast majority of human Rogue-playing experts. For example, in a test conducted at Carnegie-Mellon during a three-week period in 1983, ROG-O-MATIC had a higher median score than any of the 15 top Rogue players at the university. ROG-O-MATIC both outperforms most humans and shows some striking differences from them in its style of play. According to Mauldin, the program is careful and unimaginative. It explores efficiently and avoids all the redundant searches that humans are likely to undertake. Its fighting style, however, is rather methodical and, unless it is lucky, it misses some of the swashbuckling possibilities of human play.

Consider, for example, the factor of luck in its history-making performance last February at the University of Texas at Austin. After overcoming the incredible dangers of the deepest levels of the Dungeons of Doom, the Rogue (under ROG-O-MATIC direction, of course) found the Amulet of Yendor in a passageway on the 26th level. Racing toward the surface with his prize, the Rogue encountered on the 22nd level one of the worst dungeon monsters, a fire-breathing dragon. The Rogue drew his sword but the dragon exhaled a bolt of fire first. The searing flame missed the Rogue, who was standing in a doorway at the time, struck a wall and bounced directly back at the dragon, scorching it severely. With its remaining strength the dragon advanced toward the Rogue, who dispatched it with a sword blow. The Rogue then continued on to the surface, dealing with minor monsters as he went. When he emerged into the light of day, he had in his possession not only the amulet but also 6,913 pieces of gold and other objects.

The reader, if intrigued by the fore-



Expert subsystems and knowledge sources used by the program ROGUE going account, may want to sample the fun. For that purpose the following puzzle was constructed by Mauldin. In addition to being an introduction to the joys of playing Rogue, it illustrates the difficulty of building intelligence into the ROG-O-MATIC program.

A chamber on the 26th level of the Dungeons of Doom is currently inhabited by the Purple Worm (*P*), the Griffin (*G*) and the Rogue (@). The Amulet of Yendor (,) lies on the far side of a stairway (%) with respect to the Rogue.

```
 -----------------
|.................|
|..P.......,......|
|............%.G..|
|.............@...+
 -----------------
```

The Rogue's goal is to grab the amulet and race up the stairway without getting killed. It happens that the Rogue has only one hit point left and cannot risk a confrontation with either the Purple Worm or the Griffin. Fortunately the Purple Worm is fast asleep and the Rogue, wearing the Ring of Stealth, has only to avoid stepping on it in order not to wake it up. The Griffin, on the other hand, is very much awake and in pursuit of the Rogue. Although the Griffin would never disturb a fellow monster, it would like very much to occupy the same point as the Rogue, whom it would dismember with one slash of its cruel claws.

It is the Rogue's turn to move. Each move by either the Rogue or the Griffin involves occupying any of the eight adjacent points. So single-minded is the Griffin that it always chooses a move on the most direct path to its prey. If the Rogue makes it to the stairway, he is safe from both the Griffin and the Purple Worm: monster union rules forbid pursuit to other levels.

Last November's article on the Tower of Hanoi and the Chinese rings drew a variety of responses ranging from the mathematical to the metaphysical. Simple, nonrecursive solutions to the puzzles were presented by several readers, including Edward T. Price of Eugene, Ore. He suggests that Tower disks be painted alternately in two colors, say black and white, in order of increasing size. The puzzle is then solved quickly by adding the following rule to the ones originally given: Never place a disk on another disk of the same color. There is then no choice about where each one should be placed. The corresponding puzzle posed for the Chinese rings was solved by Morris S. Samberg of Howard Beach, N.Y. One alternates between moving (slipping on or off) the first ring and moving some other ring. Only one other ring may ever be moved. If the number of rings is odd, start with the first ring; otherwise start with the other one. Rob Hardy of Dayton, Ohio, summarizes this simple solution in the following verse:

An iterative solution
That's sure to spoil the fun:
Alternate changing the end ring
With changing some other one.

No reader was able to find a simple mathematical scheme behind the King Wen ordering of *I Ching* hexagrams. Many readers observed that hexagrams are paired across columns in a simple way, but this hardly explains the overall order. Homer E. Brown, an electrical-engineering consultant in Cary, N.C., produced a rather suggestive analysis, however: From each hexagram count up or down 10 hexagrams, skipping from the end of one column to the beginning of the next if necessary. The hexagram thus arrived at always has a simple relation to the one started with. The rule is not deterministic, however, as one never knows whether to count up or down.

In spite of this seeming relation (which has a number of exceptions in any event), I incline to the view that King Wen's arrangement follows metaphysical principles. Our attempt to project current preoccupations with science and technology into past cultures results in a distorted view of what once was. As a general rule the "scientific" systems that once existed served a role strictly subservient to a largely religious world view. Similar opinions are offered by Bernard X. Bovasso of Saugerties, N.Y. The *I Ching* is concerned with the ordering of time, maintains Bovasso, and so was King Wen. David White, a philosopher at Macalester College in St. Paul, Minn., referred to a translation of the *I Ching* by James Legge [see "Bibliography," page 128]. An appendix to this work lays out a metaphysic governing the order of all 64 hexagrams. For example, the first three are connected respectively with the concepts of heaven, earth and chaos. The chaos hexagram, denoting what is regarded as the disorder of all created things taken collectively, follows heaven and earth since all created things fill the space between heaven and earth.

Readers attempting to solve the 9-by-9 pursuit problem in the December column on sharks and fish may have become enmeshed in unnecessary complications. The ocean grid containing the two sharks and three fish should have been 9 by 10, therefore add one more vertical line to the grid.

BOOKS

Sacrificial treasures, a reflective chemist, a Quaternary indictment, planetary rings

by Philip Morrison

CENOTE OF SACRIFICE: MAYA TREASURES FROM THE SACRED WELL AT CHICHÉN ITZÁ, edited by Clemency Chase Coggins and Orrin C. Shane III. University of Texas Press ($35; paperbound, $24.50). Just 100 years ago a young "aspiring antiquarian" named Edward H. Thompson arrived as United States consul in Mérida, capital of Yucatán. Influential Massachusetts sponsors had secured him the post, although neither he nor they were much concerned with diplomatic chores. Within a decade he gave up the office and bought with his sponsor's funds the famous ruins of Chichén Itzá. Thompson then settled not far from the site, which is now the most widely visited of all the restored cities of the pre-Columbian Maya.

That city, whose name is construed as Mouth of the Well of the Itzá, was never lost to human memory. It was a holy ruin when the Spanish came and remained a center of covert worship for the countryfolk. The main road across Yucatán from Mérida passes as it always has the pyramids and columns, which stand in plain sight across the limestone flats. Yucatán is rainy in a normal summer, plenty of water for the maize and beans. But there is almost no running water there. For more than 100 miles the modern highway crosses not one bridge, hardly even a culvert. The porous limestone karst everywhere drains swiftly; all the water lies below. The people depend to this day on artificial storage or on large natural sinkholes in the limestone, where the soluble rock has collapsed down to the water table, often deep underground. The Yucatec word for these natural wells has become the Spanish term *cenote.* Where a cenote is, expect signs of ancient settlement.

The Cenote of Sacrifice was no civic water source like the others, except perhaps for ritual pots of holy water. Its deep pool lies a quarter of a mile from the sunny precincts of the pyramid, half-shaded in the thick low forest, encircled by sheer walls of rock rising 80 feet above the water, whose surface green algae darken. The ancient causeway that leads through the trees to the cenote was walked not in household routine but in solemn procession of sacrifice. So reported Bishop Diego de Landa in 1566: into this dark well the old rulers were used to "throwing men alive... also... a great many other things... if this country possessed gold, it would be this well that would have the greater part of it."

Thompson believed Landa, well-supported as the old book was by unwavering tradition. In 1904 Thompson first lowered a clamshell dredge from a derrick he had built on the south edge of the cenote, to grope amidst the 12 feet of yellow muck, imprisoning decayed trunks and branches and loose rocks, that covers the firm bottom of the cenote. He dredged throughout five summers and then spent one fruitless final season with a sponge-fisherman partner trying air-hose diving gear. What he gathered from the unmappable depths of the pool he sent to the Peabody Museum at Harvard or held in his own house. Archaeology, patient for a millennium, is profligate with the decades; what truth lay in the murk of the well has waited out revolution, fire, rumor, litigation and elaborated scholarship, to finally emerge into daylight within the past 30 years.

Landa was right: in 1940 the physical anthropologist Earnest A. Hooton identified 42 individuals represented in the human bones retrieved from the dark waters, half of them people under 20, including many children. They must have been sacrifices, prisoners or volunteers, dispatched perhaps to seek the boon of rain.

This book is an illustrated catalogue of a wide sample of the treasure of the cenote. It is based on years of study by modern scholars; never before has so much of the yield of the depths been placed on exhibit, although the old collection has for some time been shared with public museums in Mexico. A couple of hundred objects are shown beautifully photographed by Hillel Burger, many in color. Each object is closely described, and the more important pieces are accompanied by an interpretative and comparative essay by the editors.

The shifting of debris at the bottom of the cenote made impossible ordered recovery of the objects cast reverently into the well during 10 centuries. Here they have been arranged chronologically, we may be sure not always correctly, by comparison with datable artifacts found elsewhere. Many of the older finds had been "killed" (intentionally damaged) before submission to the well: jade smashed, cast gold crushed or melted, metal sheets cut, rubber and copal burned.

The most striking single object is perhaps a censer, used for the burning of rubber and copal. The object is the cut and decorated skull of a young male, "recovering from the childhood anemia so common among peoples dependent on a diet of corn after weaning." Blue-eyed Mr. Thompson was delighted at the recovery of this piece, which had wood eye disks painted with an indigo-based paint.

The late phase of the water-held treasure extends from the 13th century to the 16th, a time when the city was no longer a great capital. It was served then by more parochial artisans, its elite far from wealthy, its offerings expressive of piety, not grandeur.

The earlier period ran from the ninth century to the 11th. A superb three-inch head with jaguar headdress, carved jadeite from a distant city, bears an incised date; its delicate portraiture is masterful. A pictorial disk of embossed sheet gold is only one of a number of splendid gold objects from the early phase. A century or two then passed in which few offerings to the well were made; the period perhaps marked a near-abandonment of the city after the close of its early dominance over all Yucatán. The two phases are so distinct as to support the implicit dating; it must have taken some time for such a change to happen.

Several well-written summaries by the editors and their colleagues tell the still uncertain history of the rise and decline of the city, context for the evidence from the cenote. The early years of power demonstrate the pervading influence of an exotic military elite, the Toltecs from central Mexico. No Toltec site is as impressive as this lowland city, where they were foreigners; the art styles they favored may have developed best here far from their home. By the later phase power had flowed to another capital; the sacred well, now old in use, became more the goal of religious pilgrimage than the votive theater of a proud warrior caste.

The earlier period is all jade and

gold, cast and embossed sheet; the later period is mainly cast copper bells, wood carvings, pottery containing burned copal and rubber. In both eras the precious materials—jade, gold, copper—were the fruit of far-flung trade, possibly at times brought by the distant makers themselves. Pendants depicting frogs and monkeys are rendered in the lovely cast gold and silver alloys of Panama; there is neither metal nor hard stone in the almost universal limestone of Yucatán. From the designs we can infer that the sheet-gold objects were as a rule worked by Mayan artisans from imported metal. Even the little cast copper bells so common in the later period were imported. At the time of Columbus a group of Mexicanized Maya, perhaps kin to the Itza, had made the bells a medium of exchange in their widespread coastal trade.

The most unusual gift of the oxygen-free depths of the cenote is its preservation of many perishable organic products. A good deal of that material comes from the work of Mexican archaeologists in the 1960's. They used up-to-date scuba techniques; a suction lift fetched objects from the depths. Their underwater search succeeded where Thompson's hard-hat effort could not. Thousands of pieces of woven cotton textiles have been removed from the water in good condition, along with basketry and wood. These are still under study. The collection presented here is not an effort to span the material culture of its makers. Rather it reveals the ebb and flow of what was admired and held precious over the life of the ceremonial city.

Mayan monkey bell

Anyone who has seen the old domed building in Chichén Itzá they call El Çaracol, the snail, will marvel at one object here. It is a sacrificial knife, its flaked-chert blade hafted in resin-treated wood. The carved handle represents two intertwined snakes with reticulated skins, feathers around the eye sockets; the carving was still colored quetzal green when Thompson dredged up the knife. These paired serpents much resemble in small compass the great 50-foot entwined snakes in masonry that form the balustrade, once certainly polychromed, of the wide staircase that leads to the high viewing platform of El Caracol.

The exhibit will travel to half a dozen North American museums after its opening at the Science Museum of Minnesota; it is part of the valuable Collection-Sharing Program of the Peabody Museum.

THE PERIODIC TABLE, by Primo Levi. Translated from the Italian by Raymond Rosenthal. Schocken Books ($16.95). Neither a chemical treatise nor an autobiography, this book is a set of 20 brief chapters of lapidary recollection ordered in time. They are bound with two short science-fiction romances and an essay on the odyssey of an atom of carbon, rather resembling a piece of T. H. Huxley's that has been given a contemporary sensibility. The rare quality of Levi's ideas and their compact expression earn this volume a place on the small shelf where stand the few books that serve to successfully express the self-examination of any life lived in or close to science. Those books include H. G. Wells's *Tono-Bungay,* a couple of the novels of Hans Otto Storm and C. P. Snow, that grotesque masterpiece *Gravity's Rainbow* and very likely the covertly artful *Double Helix.* Their common matter is the daily joys and sorrows, the indelible perceptions and the shared experiences, the community, real and vicarious, the rich lode of ready metaphor, the entire emotional cast of a mind at home in the laboratory. This text exhibits at large the clarity of the Alpine foothills, although the author cannot disguise an endearing old romance with philosophy.

Dr. Levi is by the most superficial account a "factory chemist" with three or four decades of experience in the paint and varnish industry of Turin. But he is also a literary figure of high distinction, earned by two earlier memoirs. Those books had a focal spot that was burning hot: Auschwitz. Levi, chemist, poet, philosopher, acute judge of the most diverse personalities and a writer of controlled intensity, was an ill-prepared partisan in the Piedmont, thereafter an inmate and a survivor of this factory of death. The well-planned furnaces of that place are not much seen in this book, although the strong moral distillate Levi condensed during his years of survival pervades these lucid, witty and in the end loving pages.

It is more of a service to excerpt so charged a work of art than to offer a critique at second hand. Remember that each chapter of the book is given the name of a chemical element, whose presence, real or metaphorical, informs the episode.

Try an early, poignant one, about zinc. "They make tubs out of it for laundry . . . [the metal] is gray and its salts are colorless . . . not toxic, nor

Today, Northern Telecom is doing more with that code than Morse ever dreamed possible.

Based on a binary code of dots and dashes, Morse's communication system was one the whole world could use. For years, it was the primary high-speed means for the transmission of information, and for personal and business messages.

From this earliest applicat on of a digital system has emerged a new communications technology, with Northern Telecom leading the way as the world's largest manufacturer of fully digital telecommunications systems.

We're creating new, cost-effective networks that allow the communication of all forms of information—around the office, around the country, or around the world. They're responsive; they adapt to *your* requirements, not the other way around. They constantly evolve; they change with your changing needs and never become obsolete.

These Integrated Business Networks and Integrated Information Systems from Northern Telecom are fulfilling the great hope of Morse's invention: To make better, easier, efficient communications possible for everyone.

nt northern telecom

For further information write:
Northern Telecom Inc., Public Relations Dept., 259 Cumberland Bend, Nashville, TN 37228. Or call (615) 256-5900, Ext. 4264.

MORTON GROVE, IL • NASHVILLE, TN • RALEIGH, NC • RICHARDSON, TX • SAN DIEGO, CA • SANTA CLARA, CA • WEST PALM BEACH, FL

does it produce striking chromatic reactions; in short, it is a boring metal." There is one saving detail. The metal, so yielding to acid, obstinately resists attack whenever it is sufficiently pure. Which philosophical conclusion is a thoughtful young man supposed to draw? "The praise of purity...I discarded...disgustingly moralistic.... The praise of impurity, which gives rise to changes, in other words, to life ...I lingered to consider.... For the wheel to turn...diversity, the grain of salt and mustard are needed: Fascism does not want them, forbids them...; it wants everybody to be the same, and you are not." Linked to his classmates by the "small zinc bridge," the shy student preparing zinc sulfate, knowing himself an impurity of the new regime, becomes "incomparably audacious." He walks home, arm in arm with a girl for the first time, with modest, pale Rita, who cooked the selfsame dish of acid in the same hood. She was no "grain of mustard" but a reserved young woman, for whom "the university was not at all the temple of Knowledge...[but] a difficult path which led to a degree, a job, and regular pay."

Consider nickel. Levi was unemployed, a new graduate but half-proscribed. He is engaged to work in an isolated asbestos mine in the mountains where a scheme has been hatched to extract nickel, precious in wartime, from the ample tailings. "I fell in love with my work from the very first day...nothing more at that stage than quantitative analysis of rock samples: attack with hydrofluoric acid, down comes iron with ammonia, down comes nickel (how little! a pinch of red sediment) with dimethylgloxyime... always the same, every blessed day.... But stimulating and new was another sensation: the sample to be analyzed was no longer... a materialized quiz: it was a piece of rock, the earth's entrail, torn... by the explosive's force; and on the basis of the daily data...little by little was born a map, the portrait of the subterranean veins. For the first time after seventeen years of schoolwork, of Greek verbs and the history of the Peloponnesian War, the things I had learned were beginning to be useful to me."

Try a comic sample, called after nitrogen. The time is the hardscrabble postwar years in Turin, Levi and his partner are eking out a living as manufacturing chemists, precariously set up in father's apartment. "Four resistors of 1,000 watts connected illegally upstream to the meter heat their reactor." A tough cash customer comes for a few kilos of alloxan, a red true stain suited for a new line of indelible lipstick, a compound hard to find in the market.

The library of the Chemical Institute is visited, by no means a simple matter. Alloxan is there in the books. "Here is its portrait.... It is a pretty structure, isn't it? It makes you think of something solid, stable, well linked.... From the *Zentralblatt* I ricocheted to *Beilstein,* an equally monumental encyclopedia.... The sole accessible preparation was the oldest...an oxidizing demolition of uric acid." Rare in the excreta of human and mammal, uric acid is 50 percent of bird waste. These creatures are frugal of water, able to dispose of their waste nitrogen not in dilute solution but in solid form. Gold from dung, cosmetic from excrement: the task "warmed my heart like a return to the origins, when alchemists extracted phosphorus from urine." The starting material was not cheap; the canny farmers had long sold their chicken dung for fertilizer at a high price, even if it had been gleaned by the hard work of the customer, and even if it was no fit raw material for the chemist—a crude mix of dung, earth, stone and chicken lice. Home a kilo came anyway on the bicycle carrier rack.

But in Turin an exhibition of snakes had just opened. Reptiles void pellets that are 90 percent uric acid. What an idea! "Out of the question, not even a gram," the exhibitors replied. Python pellets are very scanty, worth their weight in gold and long pledged by contract to the pharmaceutical companies. A few days yielded no alloxan from the well-sifted chicken dung either. "All I got were foul vapors, boredom, humiliation, and a black and murky liquid.... The shit remained shit, and the alloxan and its resonant name remained a resonant name.... Best to return among the colorless but safe schemes of inorganic chemistry."

End back in university days, with a chapter called iron after a young student, blacksmiths' kin, who seemed made of iron. (He became the first man of one Piedmontese Military Command to be killed fighting in the Resistance.) The class they shared was the classical qualitative analysis, with a challenging unknown sample daily. "The previous lab, where I had tackled zinc, seemed an infantile exercise to us now, similar to when as children we had played at cooking.... Not here: here the affair had turned serious, the confrontation with Mother-Matter, our hostile mother, was tougher and closer. At two in the afternoon, Professor D...handed each of us precisely one gram of a certain powder.... Report in writing, like a police report... doubts and hesitations were not admissible: it was each time a choice, a deliberation, a mature and responsible undertaking, for which Fascism had not prepared us, and from which emanated a good smell, dry and clean."

The book "is—or would have liked to be—a micro-history, the history of a trade and its defeats, victories, and miseries, such as everyone wants to tell" who has reached the age when "art ceases to be long." It became even more, a gift to those, young or old, who share in the legacy of man-the-maker and seek the boon of self-understanding. High praise is due the translator: not one phrase rings false.

QUATERNARY EXTINCTIONS: A PREHISTORIC REVOLUTION, edited by Paul S. Martin and Richard G. Klein. The University of Arizona Press ($65). Who killed the Great Auk? We know the names of the two fishermen who killed the last pair known, and destroyed their single egg, at Eldey Rock off Iceland in 1844. Who killed the giant moa? Moa bones by the wagonload and moa eggshells by the acre have been found in the middens and earth ovens of the Maori voyagers who entered the South Island of New Zealand. A hundred sites are dated by radiocarbon from about A.D. 1000 to about A.D. 1500. There is no sign of major climatic change during the protracted vanishing of the diurnal forest-fringe birds. The number measurably slaughtered over time is an order of magnitude greater than the best estimate of the available population. A New Zealand popular song sums it up: "No moa, no moa, / In old Ao-tearoa. / Can't get 'em. / They've et 'em; / They've gone and there aint no moa!" The scholars agree, and they add a word or two about deforestation by set fires and a footnote acknowledging the work of the dogs and rats that came as passengers and stowaways in the big canoes.

This volume of lively and expert papers by 50 field and laboratory scientists from around the world is more or less a debate about a bold generalization: that the wave of extinctions of many large land animals and birds so conspicuous in the geologically recent past is the bloody work of man the hunter. The concept of Pleistocene overkill is not new, but it was put forward afresh by the senior editor, who had made it the basis of a similar volume of argument in 1967. The indictment is strong, although the true bill has its defects; the case is subtly shifting. There is no clear verdict yet; the jury remains hung.

The prosecution's case is put impressively in one typical graph that shows the sharp decline in the North American mammoth population. The rela-

flawless ceramic.

these powders. Again, the similarity to metals is apparent.

In metals, average properties tend to cluster in a rather tight locus. This had not been true of ceramics until recently. But now, GTE is able to consolidate ceramics with uniform microstructure, and with Weibull statistics in the range of superalloys.

Silicon nitride matrix composite cutting tools.

Grain-boundary science.

As with powder metallurgy, parts are molded and sintered. This demands special knowledge of densification additives to help create precision-molded net shapes, and provide the necessary strength in the grain-boundary phase.

GTE has developed families of additives to provide optimized properties for specific products.

Reliable "crockery"?

GTE research is pushing back the technical barriers to the use of ceramics in high-stress environments.

Ceramic matrix composites toughened with discrete particles or fibers are one thrust of the GTE search for more reliable ceramics.

Fully effective, predictable structural components are already proving to be practical in a fast-growing array of applications.

More materials science.

Materials research moves ahead in other areas at GTE, too. High-purity GaAs crystals for semiconductors; amorphous metals for brazing foils; better phosphors for ultraviolet radiation conversion—these are among the many activities currently underway at GTE.

In the box at the right is a list of some pertinent papers GTE people have published relating to research in advanced ceramics. For any of these, you are invited to write GTE Marketing Services Center, Department TP-M, 70 Empire Drive, West Seneca, NY 14224. Or call 1-800-828-7280 (in N.Y. State 1-800-462-1075).

Pertinent Papers

Phase Effects in Silicon Nitride Containing Y_2O_3 or CeO_2: Part I—Strength; Part II—Oxidation. J. American Ceramic Soc., May, 1980.

Microstructural Effects Influencing Strength of Sintered Silicon Nitride. International Conference on Ultrastructure Processing of Ceramics, Glasses and Composites, February, 1983.

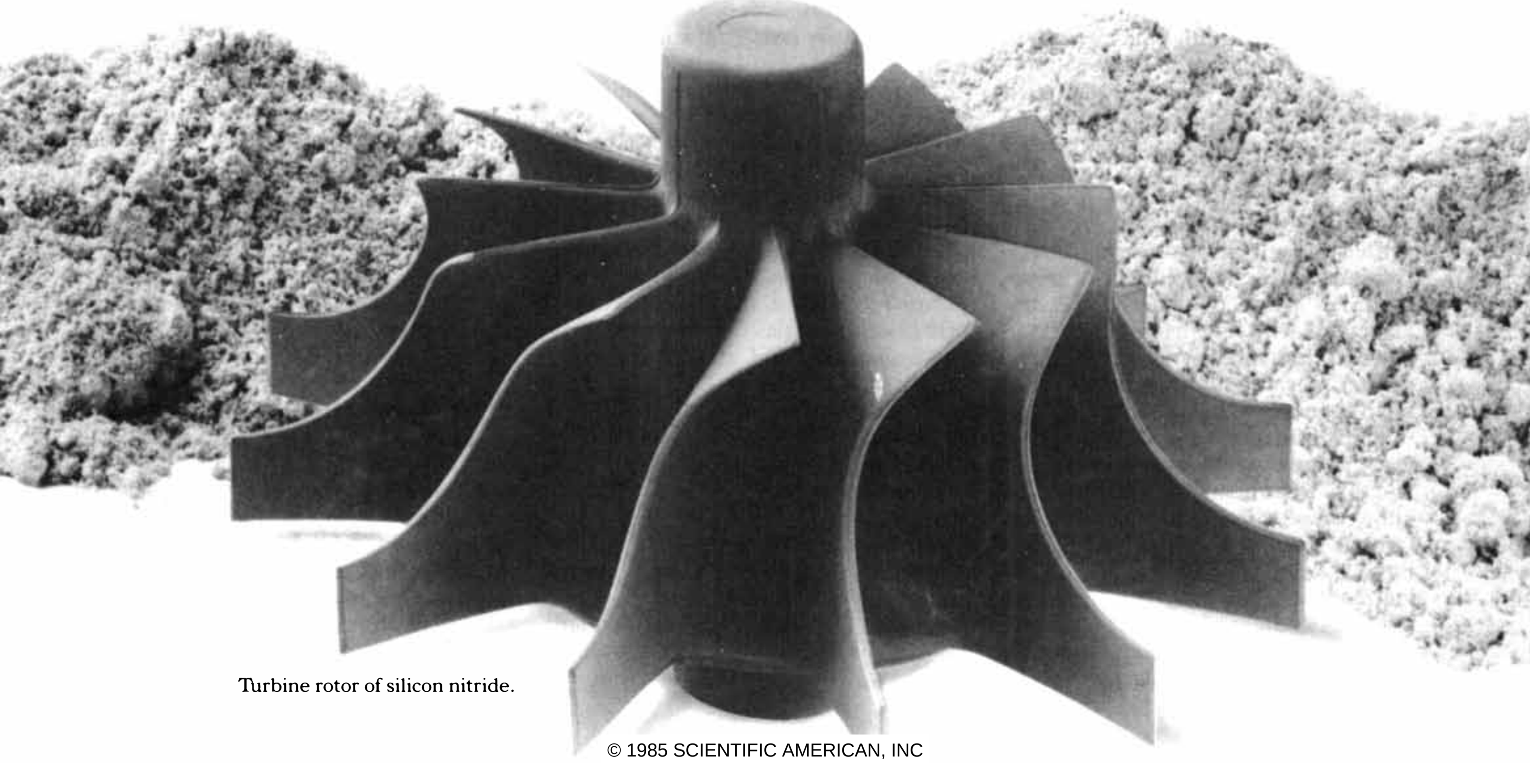
Fabrication of Turbine Components and Properties of Sintered Silicon Nitride. ASME 82-GT-252, 1982.

Advanced Silicon Nitride Based Ceramic for Cutting Tools. SME MR-83-189, September, 1984.

Design and Wear Resistance of Silicon Nitride Based Composites. Second Int'l Conference on Science of Hardened Material, 1984.

Controlled Transient Solid Second-Phase Sintering of Yttria. J. American Ceramic Soc., 1981.

GTE



Turbine rotor of silicon nitride.

tive number of elephants—all known sites are mapped and dated to form the estimate—suddenly declines from a long-occupied plateau to a tenth of that number between 13,000 and 11,000 years before the present. The North American population of *Homo sapiens* shows its first sharp rise at just that time.

The indictment is worldwide. In Africa the Plains of Serengeti are current demonstration of what a world of large browsers and grazers and their big predators must have been like, when not so long ago bison, elephant and lion alike roamed France and the Dakotas. But African game have had a long time to adapt to the primate hunters, to their fire, to their developing tricks and tools. Ten times as many genera of large mammals have vanished from the Americas within the past 15,000 years as have vanished from Africa in 100,000 years. The islands tell their own special tales: join to the moas the big lemurs and the elephant birds—the rocs—of Madagascar and the cow-size marsupial browsers of Australia, all gone once humans arrived, and the courtroom grows still. The first sudden coming of the hunters spelled extinction to the biggest of their game.

The defense is eloquent too. Are those dramatic census changes real? Is the dating right? Is the time coincidence truly causal? If it is, how can a few mammoth hunters' butchering sites match the moa hunters' ubiquitous boneyards? The prosecution reminds us that if the hunt were suddenly begun and soon over, a blitzkrieg, then the total take would be small and hard to find. It will not escape a thoughtful juror that here the prosecutor has neatly found a way around the absence of the surest evidence, the corpse itself.

The ecologists weigh in with their diverse evidence, counting pollen and beetle species, tracing drainage regions and floral associations, looking at how plants change (herbivores feed on them at the grass roots) as the lands drifted toward aridity while the glaciers dwindled. The genetics of populations is drawn on: extinction seems associated with dwarfism. What forces selected for smaller individuals and shorter gestation periods? Could it have been hunting pressure?

A set of exemplary sites are given very interesting, knowledgeable and lengthy treatment by authors who know them well. An overview of the famous mammoth remains of Siberia is the kernel of an authoritative Soviet contribution. Here are structures made of mammoth bones, all right, but there are even larger graveyards that were certainly accumulated by long natural alluvial processes. The La Brea tar pits on Wilshire Boulevard are reviewed, with the valuable reminder that the scene at that live-baited natural trap was not quite as dramatic as some people imagine. Enmirement of a single creature every few years was plenty in due time to fill the museum cases. Caves in the lower Grand Canyon were long home to the big browsing Shasta ground sloth. Its well-preserved dung has been studied for a reading of the menu, and now that glimpse of the climate by way of the desert flora has been elaborated by study of the fossil middens of pack rats in the same locality. The ground sloth disappeared at "a time... which should have been nearly ideal for its continued existence." Hunters? No sites show them.

A couple of up-to-date bestiaries add concreteness. There were noble beasts in those days: consider a European rhino with a single six-foot horn, or a massive African antlered giraffe, neck unelongated. The Australian zoo of the past is vividly illustrated. The biggest marsupial was a strange browser with a clumsy-looking development of the upper lip, perhaps an incipient trunk? It is not hard to extrapolate and envision the coming of a fine marsupial mammoth.

The experts from down under are not impressed by the case against the hunters. All we know of the economic system of the aboriginals does not provide for hunting large game. Nowhere in Australia does the archaeology yet support overkill, and it is not hard to spell out how a climate turning dry increases the space between water holes beyond tolerable distances. In a desert world, megafauna depend on free water; effectively adapted small mammals do not.

There are three distinct summations, all full of interest, and an opening excursion into the history of the idea. What makes a verdict particularly hard to reach is the plain fact that the hunters too form part of the same climate-sensitive fauna. They could emigrate to the Americas in the first place, by way of the transient land bridge of Beringia, only as the result of "a unique set of climatic events." "Do I detect a conspiracy?" asks one judge. The thick volume is uncommonly readable and varied for watchers of paleontology and the rise of humankind.

RINGS: DISCOVERIES FROM GALILEO TO VOYAGER, by James Elliot and Richard Kerr. The MIT Press ($17.50). PLANETARY RINGS, edited by Richard Greenberg and André Brahic, assisted by Mildred Shapley Matthews. The University of Arizona Press ($35). Galileo saw something unique about Saturn, twin big satellites perhaps, but his optics did not provide a clear view. His interpretation was naturally Copernican: a three-bodied planet. Within 50 years there were competing theories. A fine model simulation was run by the Accadèmia del Cimento under the patronage of Prince Leopold de' Medici. The experts and an untrained control group gathered to view lighted models set up at the end of a long gallery; each group decisively favored Christiaan Huygens' Cartesian solid ring around Saturn rather than a complicated satellite model.

Both these books celebrate the new era: two, three, many rings. The first book is a personal and knowing account by a participant working with a veteran science writer. The second is a rich symposium, a score of papers at the research level. Full of mathematics, they form an array more coherent than most such proceedings. There are useful illustrations in both (the smaller Elliot-Kerr book draws a good deal on the fuller report); the big volume ends with a dozen color plates of Saturn's rings in many aspects, along with computer-enhanced color, simulated images and even a painting by the artist-astronomer W. K. Hartmann of what Saturn's rings might look like from a few hundred yards above their plane.

The point is of course that between 1976 and 1983 rings were unexpectedly found around Jupiter and Uranus. (Neptune has shown no ring for sure but cannot yet be certified ring-free.) Then the rings of Jupiter and later of Saturn were superbly imaged and examined in detail during close visits by the two Voyager probes.

It was Elliot who led the discovery of the rings of Uranus. That find was made possible by the flowering of contemporary technology. Just as Galileo's first half-recognition rested on a powerful new development that was paramedical (spectaclemaking), so is our astronomy often paramilitary.

The rings of Uranus appeared first as spiky dips in recorder charts that plotted second by second the output of infrared detectors. From the stratosphere a remarkable NASA jet-aircraft-borne telescope kept the moving planet under watch during a tightly scheduled rendezvous with a long-predicted star shadow path. The plane was there on time 1,000 miles east of the forbidding island of Kerguelen in the Indian Ocean, and eight miles above the midnight sea.

Two chapters of Elliot's book tell a charming tale of the discovery of the rings of Uranus, from proposal to pub-

A close look at Saturn's rings

lication, with all the vicissitudes and pleasures of the new style of astronomy. What the Cornell investigators were after was a quick look through the atmosphere of Uranus at the predicted time and place of the rare eclipse of a modest star by that planet. The color and intensity of the light as total eclipse approached and receded probed the planet's atmosphere. It was the symmetry of the flicker pattern, once the charts were spread out on the floor after the return to Ithaca, that disclosed the rings. The team had made a joke of their experiment: if it did nothing else, it might set limits to unknown and unexpected rings. Instead it showed them plain as plain could be, and by now nine have been mapped in detail in a dozen more flights since 1977, Uranus eclipsing fainter and more commonplace stars.

What kinds of rings do we list now? There is a pet shop full, if not yet a zoo. Jupiter has "ethereal" rings, like the outermost ones of Saturn, wide and diaphanous. Uranus has discrete dull-black rings, narrow as wires, some eccentric and even coarse-textured. Saturn has bright broad rings and gaps superficially smooth that at second glance break into some 1,000 narrow bands. A few narrow rings are there; what look like braided rings are probably only kinks. No one doubts that all we see is a slowly changing array of chunks of orbiting material, fine dust to little moonlets. No one has yet seen a ring particle, but it is clear that rings of solid or fluid matter would be swiftly torn apart by the tidal forces.

The ingredients of the explanations for all of this are not hard to grasp, although the dynamical cookery is subtle and our best recipes are incomplete. Begin with many-body gravitational interactions, the ordinary satellites on their rounds finding harmonic resonances in the orbits of the myriad ring particles. Here is a source of gaps and even of a little piling up.

Add the recognition that the ring particles themselves act as a gravitating sheet, capable of supporting traveling spiral wave patterns of variable density by their mutual attraction. Do not forget the presence right in among the ring particles of ring moons, giants of the ring population, yet often too small to perceive in the images. One pair of such moonlets can be seen in the act of confining a stream of particles into a thin dense thread. Such a gravitational shepherding process is by no means simply teased out of the Newtonian interactions; it is there if the theorists are allowed some dissipative mechanism on the side.

The processes involved go beyond gravitation, although "most people don't realize how many solutions" Newton's equations have, as one of the theorists puts it. The finest ring particles respond more to electromagnetic and radiative forces than to gravitational ones. The faint spokes that slowly drift around the rings of Saturn find explanation in that way. They consist of small, highly visible populations of charged dust levitated above the ring plane, marching to a quite different dynamical drummer.

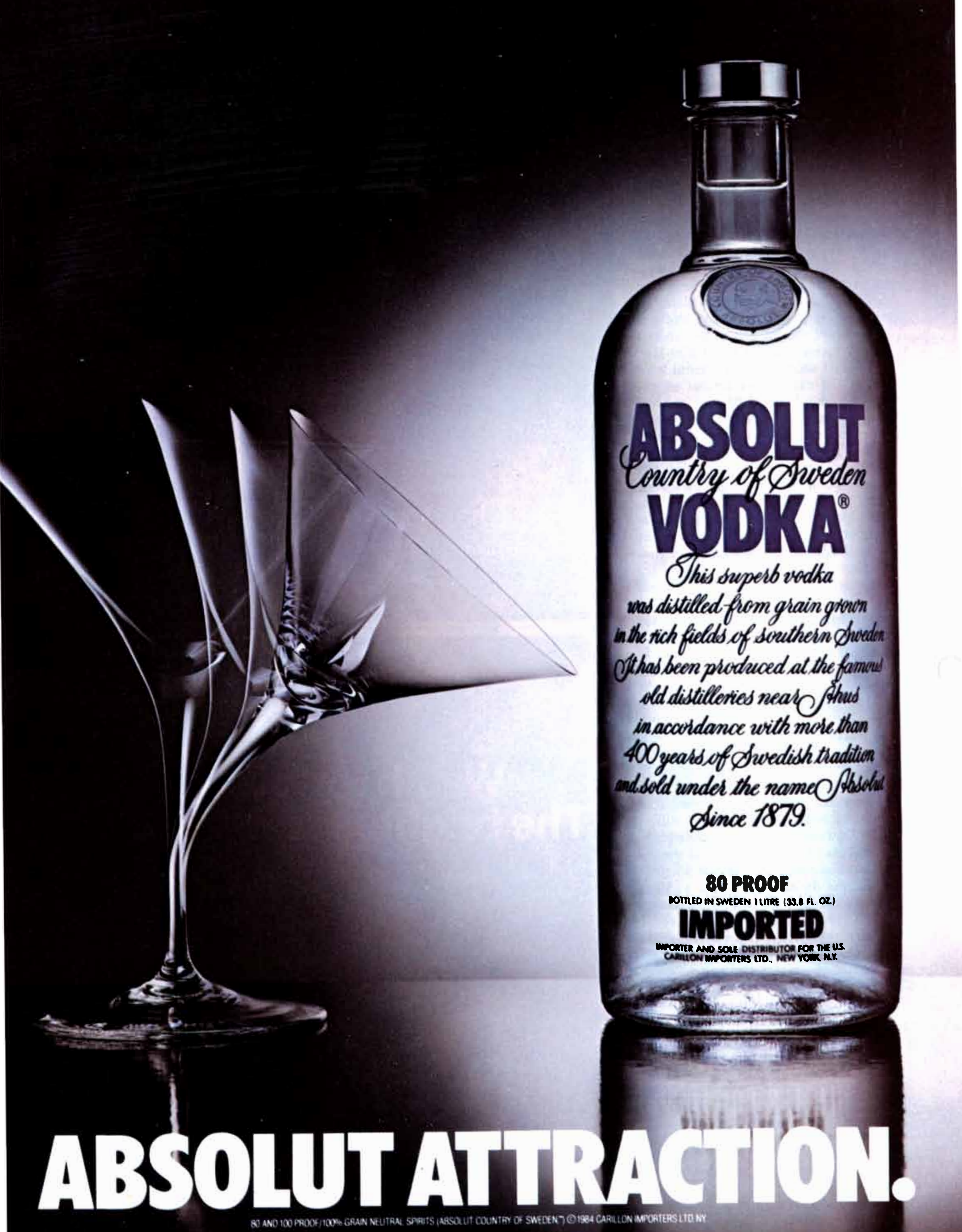
Moonlets are sources of erosional dust and may hide in narrow dust rings that they make and hold by a strange orbital dance, first being overtaken by and then overtaking their progeny. Throughout the entire pageant particles large and small collide. Masses both aggregate and disaggregate everywhere; the stickiness of the bouncing snowballs and the disrupting tidal forces of the planet seek equilibrium.

Out of such parts models are built to try to fit the copious data. The excitement of the community of planetary scientists is apparent. These shifts of opinion are the burden of the unfolding narrative that is the Elliot-Kerr book. The second, more formal book makes the excitement almost as evident through learned papers that propose alternative explanations for the complex and often ambiguous data.

The rings evolve. Fine particles will not last forever in orbit; they are cleansed out of the system in several ways. We have to explain rings for the most part as recurrent phenomena, more or less in a steady state, unless we are willing to claim the great good luck of catching a fleeting pattern in the long history of a giant planet. Some recycling is believable enough, but origins and lifetimes must be understood to complete the success of any model for what we see. Both the disruption of satellites now gone and the relict ring left over from the formation of the central planet are plausible origins.

Voyager is on the way to Uranus, due there punctually about a year from now, just the month that Halley's comet will make the headlines. With skill and luck good images will show—or eliminate—the putative shepherd moons in the thin rings. Other satellites will follow, and Space Telescope Hubble too can take a good look at spokes and distant rings. In 1989, if all goes well in the decade-old circuitry, Voyager may reach Neptune able to check sensitively for rings so far obscure.

The rings emerge from these books as part of our solar-system home. They are structures as familiar as the Grand Canyon of the Colorado or the rays of Tycho. In rings a myriad icy boulders in swift flight clash silently on a scale far greater than that of the wave-stirred stones on Dover Beach. It would be a pleasure to learn why some rings are snow white and some coal black; that is not in the books yet, but we can expect it someday, with a long list of other problems still not clear. Meanwhile, at the level of the general reader as well as that of the research worker, there is plenty to learn; these books offer two fine places to start.


ABSOLUT
Country of Sweden
VODKA®
This superb vodka
was distilled from grain grown
in the rich fields of southern Sweden.
It has been produced at the famous
old distilleries near Åhus
in accordance with more than
400 years of Swedish tradition
and sold under the name Absolut
Since 1879.
80 PROOF
BOTTLED IN SWEDEN 1 LITRE (33,8 FL. OZ.)
IMPORTED
IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.
CARILLON IMPORTERS LTD., NEW YORK, N.Y.
ABSOLUT ATTRACTION.
80 AND 100 PROOF/100% GRAIN NEUTRAL SPIRITS (ABSOLUT COUNTRY OF SWEDEN™) © 1984 CARILLON IMPORTERS LTD. NY.

Established 1845 SCIENTIFIC AMERICAN February 1985 Volume 252 Number 2

Predicting the Next Great Earthquake in California

On part of the San Andreas fault an earthquake of magnitude 8 in the next 30 years is assigned a probability of 50 percent. Precise predictions await better knowledge of how quakes are triggered

by Robert L. Wesson and Robert E. Wallace

With the successful prediction of the eruption of Mount St. Helens in May, 1980, earth sciences in the U.S. entered the era of real-time geology: the study of geologic processes as they happen, sometimes minute by minute. The prediction itself was far from perfect; indeed, it lacked a short-term warning. Still, the advice of earth scientists led officials in Federal, state and local government to institute a series of precautionary measures, including the closing of forest areas, which doubtless reduced the loss of life and property caused by the eruption.

In California the geologic event most in need of successful prediction—the one whose destructiveness could most be averted—is a catastrophic earthquake. In geologic terms California is a realm of earthquakes: it is part of the Ring of Fire, the belt of earthquakes and volcanic activity that circles the Pacific. The San Andreas fault, which scars California from north to south, is one of the earth's great emblems of seismic activity. In human terms California is the most populous state in the U.S. and a center for many of the nation's critical technology-oriented industries. Some 10 percent of the nation's population and industrial resources are there, and 85 percent of these resources (or 8.5 percent of the nation's total) are in a strip of 21 counties along the continental margin that are well within the seismic domain of the San Andreas. A single one of those counties—Santa Clara County, in northern California—manufactures a fourth of the nation's semiconductors. Santa Clara County suffered heavy damage in the 1906 earthquake that devastated San Francisco.

In the wake of the Mount St. Helens eruption the National Security Council assembled a group of investigators and agencies, including the U.S. Geological Survey, to predict the consequences of a hypothetical earthquake in California. The group hypothesized several potentially damaging quakes, among them an earthquake similar to the great quake of 1857, which occurred along the southern San Andreas fault. The recurrence time for such a quake is now estimated to be about 140 years; the likelihood of such a quake's occurring sometime in the next 30 years is estimated to be about 50 percent.

According to the analysis, the hypothetical quake would cause between 3,000 and 13,000 deaths. The lower figure would apply if the quake occurred at 2:30 A.M. The reason emerges from worldwide observations of how earthquakes affect buildings like those in California. Briefly, the population of California is safest when it is home in bed. Wood-frame, single-family houses strongly resist structural damage by earthquakes. The higher figure would apply at 4:30 on a weekday afternoon, when much of the population is at work or on the streets. Again the reason emerges from architecture. Old masonry buildings unreinforced by steel are likely to collapse in even a moderate quake. During the Coalinga, Calif., quake in May, 1983, 30 out of the 40 pre-1930 masonry buildings in the town collapsed or almost collapsed. Such buildings are common in the center of virtually all the older cities and towns in California; some 8,000 of them are in Los Angeles alone. They are gradually being replaced, particularly in Los Angeles, but significant numbers of them will remain for many years.

Earthquake prediction is at the cutting edge of basic science. Substantial research programs are under way in China, Japan and the U.S.S.R. In the U.S., meanwhile, some $17 million is being spent each year on prediction and another $43 million on a research program including work on earthquake-hazard assessment, the engineering of buildings to resist earthquakes and studies of the fundamental nature of earthquakes. Progress on short-term prediction has been slower than workers hoped it would be a decade ago, when early successes were reported by both Soviet and American researchers. On the other hand, progress in long-term prediction and in the understanding of how earthquakes occur has been substantial. Now more than ever the aim is to learn to interpret the signals of strain accumulating in the earth within the framework of fundamental understanding of the earthquake process.

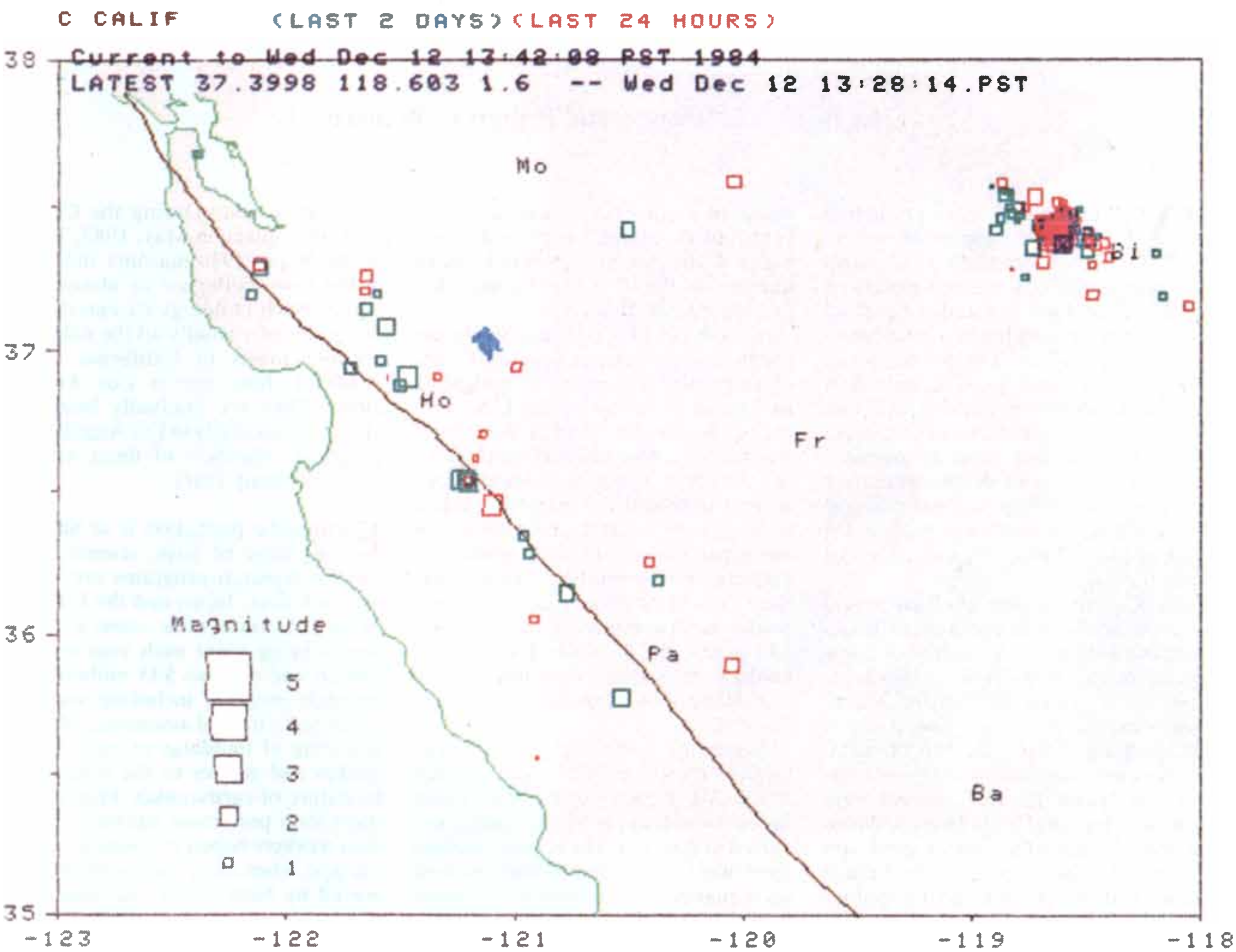
The first modern understanding came early in this century, when the

San Andreas fault was recognized as being the cause of great earthquakes; it had produced the San Francisco quake of 1906. In the wake of that catastrophe Henry F. Reid of Johns Hopkins University proposed that earthquakes are generated by the sudden slip and elastic rebound of crustal blocks bordering a fault. In essence he proposed that the crust could bend under strain, then snap and straighten out, releasing the energy of the strain.

Reid's model, now coupled with the theory of plate tectonics, which emerged in the 1960's, is the basis for estimating when and where large earthquakes can be expected. In the theory of plate tectonics most major earthquakes are explained as sudden movements or ruptures along the faults constituting the boundaries of the earth's great crustal plates. According to the theory, the San Andreas is one of a set of faults representing the boundary between the Pacific plate and the North American plate. The plates are sliding past each other; in particular the sea floor of the Pacific is moving toward the northwest with respect to North America. Over the long term the San Andreas must accommodate slippage of about 3.5 centimeters per year. The inescapable conclusion is that sudden slippages in the future will give rise to potentially catastrophic earthquakes.

How, precisely, do earthquakes fit into the framework of the plate-tectonic theory? Fundamentally the slippage between two plates proceeds without seismic disturbance at depths in the earth greater than 10 to 12 kilometers. The rocks there are hot enough to be ductile; that is, they can deform continually. (Putty, for instance, is ductile when the stuff is slowly deformed.) Closer to the surface, the rocks of the crust are brittle, so that slippage along a zone of weakness—in particular a fault—is controlled there by friction. The sudden slippage between blocks in the upper crust is what produces an earthquake. The slippage may be initiated by breakage at a particularly strong (high-friction) part of the fault, known to geologists as an asperity.

How, then, can earthquakes be predicted? A beginning to long-term prediction was made when S. A. Fedo-



SEISMIC ACTIVITY in central California during the 48 hours preceding 1:42:08 P.M. Pacific standard time on Wednesday, December 12, 1984, is displayed in this computer-generated chart. Seismographs distributed throughout the state detected the activity; they are linked to a computer at the U.S. Geological Survey in Menlo Park. The computer determined the size and location of earthquakes; its record is updated every few seconds. San Francisco is at the upper left; it lies, with Hollister (*Ho*) and Parkfield (*Pa*), along the San Andreas fault (*black line*). Modesto (*Mo*), Fresno (*Fr*) and Bakersfield (*Ba*) are farther inland. The concentration of earthquakes at the upper right is part of a burst of activity that began with a magnitude-5.7 quake on November 23, 1984, near Bishop (*Bi*). A quake of magnitude 3 can be felt by people in the vicinity; a quake of magnitude 5 can cause local damage; a quake of magnitude 7 is felt for hundreds of thousands of square miles and can cause extensive damage; a quake of magnitude 8 is considered a great quake. Each increase of 1 on the magnitude scale represents a tenfold increase in shaking and a thirtyfold increase in energy.

tov of the Institute of Physics of the Earth in Moscow and K. Mogi of the Earthquake Research Institute of the University of Tokyo plotted the rupture zones, or areas of crustal slippage, for great earthquakes along the western margin of the Pacific. The zones could be inferred from the spatial distribution of the aftershocks for each quake. Fedotov and Mogi both noted that the zones tended to abut without overlapping, and that over a long period (a century or two) the zones could be expected to fill the Pacific margin. That is, the entire plate boundary could be expected to rupture.

At Columbia University, Lynn R. Sykes, John Kelleher and William R. McCann applied these expectations to a study of earthquakes along major plate boundaries throughout the world. Several places along those boundaries were notable. Each place had been the site of a great earthquake more than a century ago. Since then, however, no earthquake at the site had represented a significant rupture of the plate boundary there. Sykes and his colleagues identified each such place as a "seismic gap": a candidate for a great earthquake in the next few decades. Several such gaps have already ruptured, fulfilling the forecasts.

The San Andreas fault in southern California is a major seismic gap. That is not to say the fault responds in the same way throughout its length to the displacement of crustal plates. Nor is it to say that the identification of seismic gaps provides the only basis for long-term earthquake prediction. On one section of the San Andreas fault in central California, fault creep, or slow slippage, characterizes the displacement. In other places the fault releases its stored elastic strain in characteristic steps no longer than a few centimeters. In still other places the characteristic slip increment is from seven to 10 meters. Episodes of slippage of the latter type were responsible for the great earthquakes of 1857 and 1906.

Averaged over a sufficiently long period of time the sum of the various slippages or displacements—the slow, aseismic fault creep, the fault displacement accompanying earthquakes, and inelastic deformation such as crustal folding—must equal the displacement between two plates. This leads quite naturally to the idea of long-term earthquake prediction based on what might be called a slip budget. In 1970 one of us (Wallace) employed Reid's elastic-rebound theory in concert with the slip rate determined from the offset of geologic features across the San Andreas fault over the past 20 million years. The result was the very-long-range prediction, more properly called a statement of earthquake potential, that large earthquakes would occur somewhere along the fault at intervals of from 50 to 200 years.

In recent years a further means of long-term prediction has arisen: new geologic techniques have created the science of paleoseismology. The techniques combine detailed studies of landforms across fault zones, analysis of deformed layers of sediment in the walls of trenches excavated across active faults and the determination of the age of carbonaceous material found in the sediments. (The last is done by radiometry: specifically the measurement of the concentrations of carbon isotopes.) In this way the techniques yield the approximate time of occurrence and the relative energy of major ancient earthquakes; thus they provide a much longer baseline of earthquake occurrence than any historical record can give.

The most complete baseline established so far has emerged at Pallet Creek, on the San Andreas fault about 55 kilometers north of Los Angeles. It is the work of Kerry E. Sieh of the California Institute of Technology. At Pallet Creek successive episodes of slippage along the San Andreas fault have repeatedly disrupted the sedimentary layers in the bed of the creek and at the same time blocked the creek's drainage, so that the disrupted layers would be covered by newly deposited sediment. In sum, the blocked drainage and repeated ruptures have formed a complex pattern of fractured and refractured layers of peat, silt and sand. In the mid-1970's Sieh made excavations there. He analyzed each displacement. The peat layers provided material for carbon-isotope dating, from which the age of each episode of faulting could be inferred.

Sieh's initial analysis revealed a history of at least nine large earthquakes in the past 1,400 years. Then, as more excavations were made, the number grew to 12. Sieh now thinks the intervals between large earthquakes along the Mojave segment of the San Andreas fault—the segment on which the Pallet Creek site is situated—range from 50 to 300 years. According to Sieh, the average recurrence interval is 140 to 150 years.

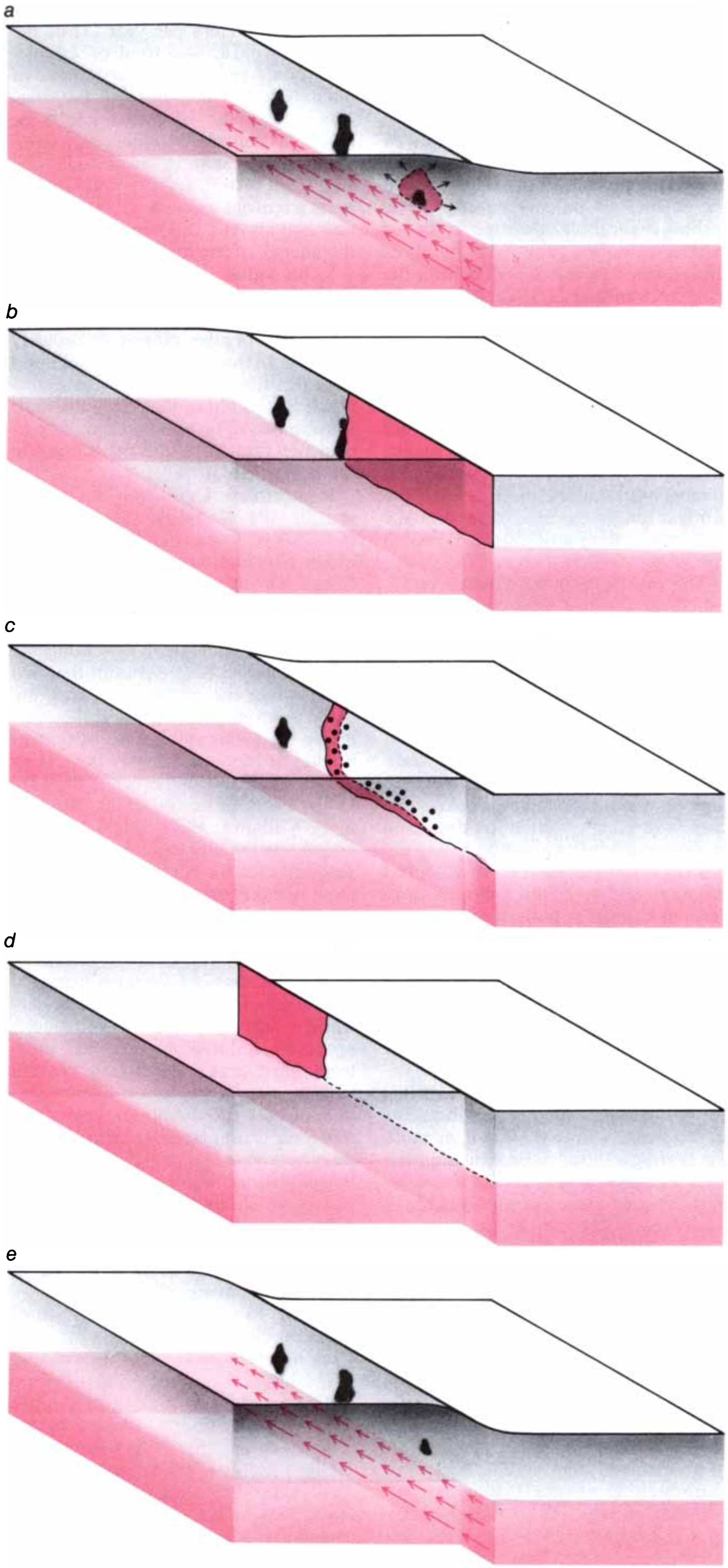
That number should be compared with the recent geologic history at the site. The last great earthquake on the Mojave segment of the San Andreas fault was the one in 1857. The number of years since then is 128, a number alarmingly close to Sieh's calculation for the average interquake interval. Moreover, the rate at which strain accumulates along the San Andreas fault is 3.5 centimeters per year. Thus the strain since 1857—a total of 4.5 meters—could generate an earthquake of magnitude 7.5 or greater at any time. (An earthquake of magnitude 5 or greater is considered potentially damaging. Each step in magnitude—the step from 5 to 6, for example—represents a tenfold increase in ground motion and about a thirtyfold increase in the energy of the quake.)

Sieh's value for the average recurrence interval was incorporated into the prediction the Geological Survey prepared for the National Security Council. In that analysis the likelihood of an earthquake of magnitude 8.3 along the southern San Andreas fault is estimated to be between 2 and 5 percent per year, or about 50 percent in the next 20 to 30 years.

In northern California no paleoseismic site along the San Andreas fault has yet been found that yields as much information as the site at Pallet Creek. Nevertheless, the available evidence suggests a recurrence interval of 150 to 300 years for an earthquake of magnitude greater than 7.5 along the part of the San Andreas fault that last ruptured in the earthquake of 1906. Fifty years of seismic quiescence followed the 1906 catastrophe, but earthquakes of moderate degree have been occurring since 1957, much as they did in the decades before 1906. Accordingly William L. Ellsworth and his colleagues at the Geological Survey suggest that over the next four decades the most likely prospect is earthquakes of magnitude 6.5 to 7, which can be highly damaging locally.

The processes of long-term strain accumulation and release are understood relatively well. In contrast, the processes holding the key to short-term earthquake prediction remain largely unknown. For example, the details of how the ductile movement of the lower crust places strain in the brittle crust above it is poorly understood. Perhaps more important, the detailed way in which the brittle crust fails and ruptures is not yet known. To what extent, for instance, is the rate at which strain accumulates in the brittle crust uniform? What significance should be attached to variations in the rate? What precisely is responsible for initiating an earthquake? Is it an external stress that acts as a trigger? Is it a surge of strain accumulation? Or is it some progressive process of failure in the crustal rocks or the fault zone itself?

One strategy is central to all the efforts at short-term prediction. Geologic phenomena must be observed in the period preceding the main earthquake shock. Some results are already avail-

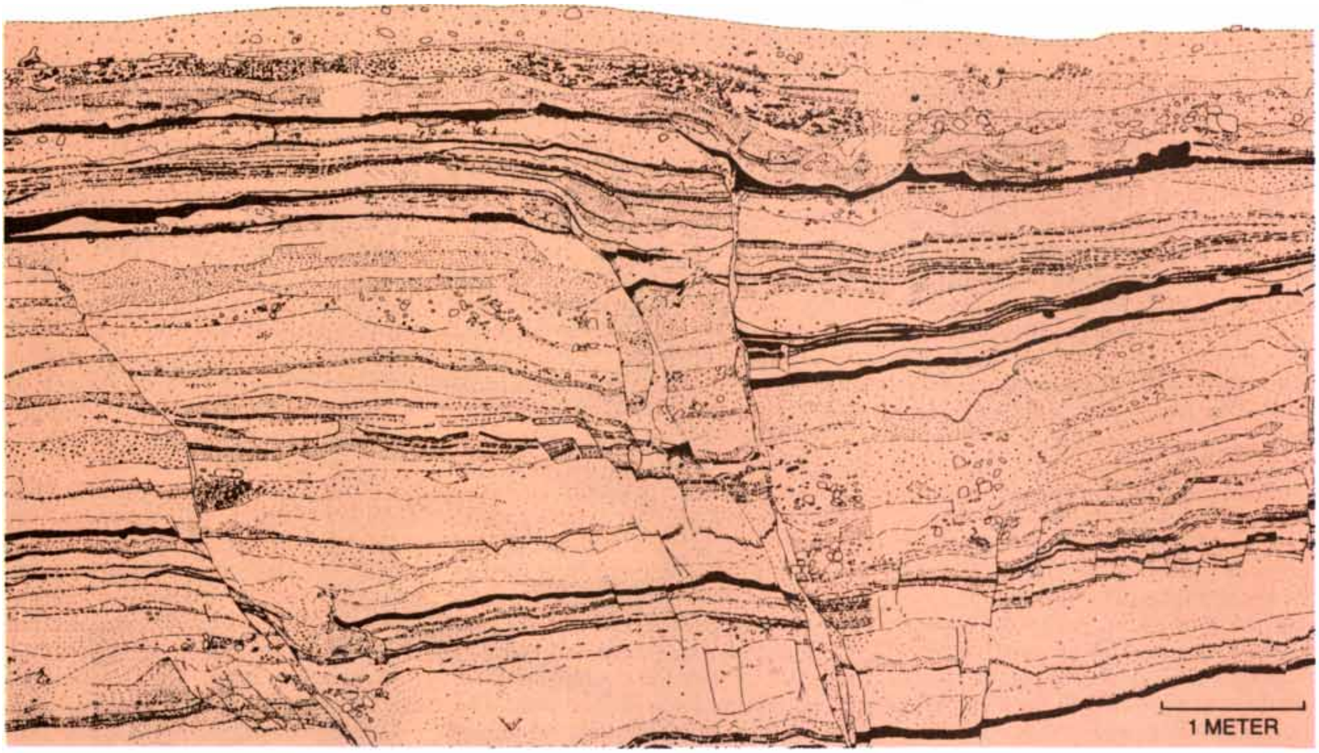


able. Prior to many large quakes the background seismicity of the region seems to increase, gradually or suddenly. The clearest instances, perhaps, were the swarms of small earthquakes preceding the 1975 Haicheng earthquake in Liaoning Province in China. The swarms began some two months before the main shock, but Chinese investigators had already decided the region had the potential for a very large earthquake. They had based their prediction on a long-term pattern of earthquake migration: the positions of quakes seemed to move along a belt several hundred kilometers long. Moreover, some geodetic measurements (literally measurements of the shape of the land) had suggested that strain was accumulating.

The site itself had experienced no more than a few very minor earthquakes for decades before the major quake. But when the earthquake swarms became vigorous on February 1, 1975, residents of unreinforced masonry homes were evacuated, so that when the main shock came, on the evening of February 4, and many houses collapsed, great numbers of lives undoubtedly were saved. In marked contrast, the Tangshan earthquake some 18 months later was not preceded by any recognized foreshocks. It caused a loss of life officially estimated at 240,000.

Seismicity is a notable portent of major earthquakes; it is not, however, the only portent. In addition, accelerated deformation of the crust is sometimes, although not always, detected before large quakes. It was detected, for instance, before an earthquake at

EARTHQUAKE CYCLE is now understood in the context of the theory of plate tectonics. Basically an earthquake results from a sudden rupture along the fault where two great crustal plates are moving past each other. Before the instant of the rupture the plates have been moving smoothly at depths greater than 10 to 12 kilometers; the crust there is sufficiently hot so that it is ductile (*colored arrows*). Meanwhile at shallower depths, where the crust is cold and thus brittle, stress has accumulated at asperities, or locked points (*black*). At the instant of rupture (*a*) an asperity breaks, so that crustal slippage (*dark color*) spreads through the upper, locked part of the fault. In 100 seconds or so the rapid slippage ends (*b*). For the next several months aftershocks (*black dots*) ensue (*c*). Slow, additional slippage at or near the surface may also occur. At some much later time, say 50 or 100 years, the adjacent section of the fault may break in a major earthquake (*d*). Finally, after perhaps two centuries, the cycle begins again: the displacement of the deep crust has imposed new stress on the upper, brittle crust (*e*).



TRENCH AT PALLET CREEK, along the San Andreas fault about 55 kilometers north of Los Angeles, was excavated by Kerry E. Sieh of the California Institute of Technology. The excavation exposed disruptions in layers of sand (*dots*), gravel (*irregular stipples*) and peat (*black*), and so revealed successive episodes of faulting associated with ancient earthquakes; the ages of the peat beds, determined by radiocarbon dating, bracket the times of the quakes. The uppermost peat bed is unbroken; it overlies the faults, which are older. Further consequences of each quake include the folding and deforming of sediments that were soft at the time of the quake. In some places sandy material along the line of a fault signals the liquefaction of the sand by the force of a quake. Sieh's work, termed paleoseismology, aids earthquake prediction by yielding earthquake statistics spanning a period far longer than any historical record.

Parkfield on the San Andreas fault in 1966. Certain laboratory experiments and theoretical models of the earthquake process suggest that accelerated deformation—called preseismic slip—is intrinsic to earthquakes. The idea is that the failure of crustal rock is preceded by the development of small cracks or weak spots, which then grow into the catastrophic rupture. The identification of earthquake precursors within the background of crustal deformation, which fluctuates continually, remains a chief objective of earthquake-prediction research.

The instrumentation required to monitor the crust and detect the anomalous events that may foreshadow an earthquake has advanced rapidly in the past decade, both in capability and in the reliability of individual devices. The instruments have also expanded in quantity, along with the geographic area they cover.

The workhorses of the effort are inexpensive, reliable seismometer packages, developed by Jerry P. Eaton and his colleagues at the Geological Survey. More than 500 such packages, each one consisting of a seismometer and supporting electronics, are now deployed throughout California. By means of telephone or radio links they continuously relay seismic signals to recording and processing centers in Menlo Park and Pasadena. In 1984 they registered more than 10,000 earthquakes. Computer hardware and software developed by Rex Allen and James O. Ellis at the Survey automatically determine the location, depth and magnitude of each quake; on some days the computers automatically cope with as many as 1,200 quakes.

The instrumentation also includes several networks of laser distance-ranging devices. Here the lengths of the sides of a network of triangles are monitored. A change in the lengths yields a measure of the crustal stress along a line; a change in the area enclosed in any one triangle yields in addition a volumetric measure of crustal strain. With appropriate corrections for the atmosphere's optical properties, employing air-temperature measurements made from an aircraft flying along the line of sight, distance measurements accurate to within about a centimeter over a line of 30 kilometers are possible. Recently some two-color laser systems have been put in service. The beams of different color have different optical properties; thus atmospheric variations can be calculated directly from the difference in the transmission time for the beams. In that way atmospheric variations can be accommodated automatically. Vertical strain in the crust corresponds to changes in the elevation of the land, as measured directly by standard surveying techniques and indirectly by determining the local strength of the earth's gravitational field.

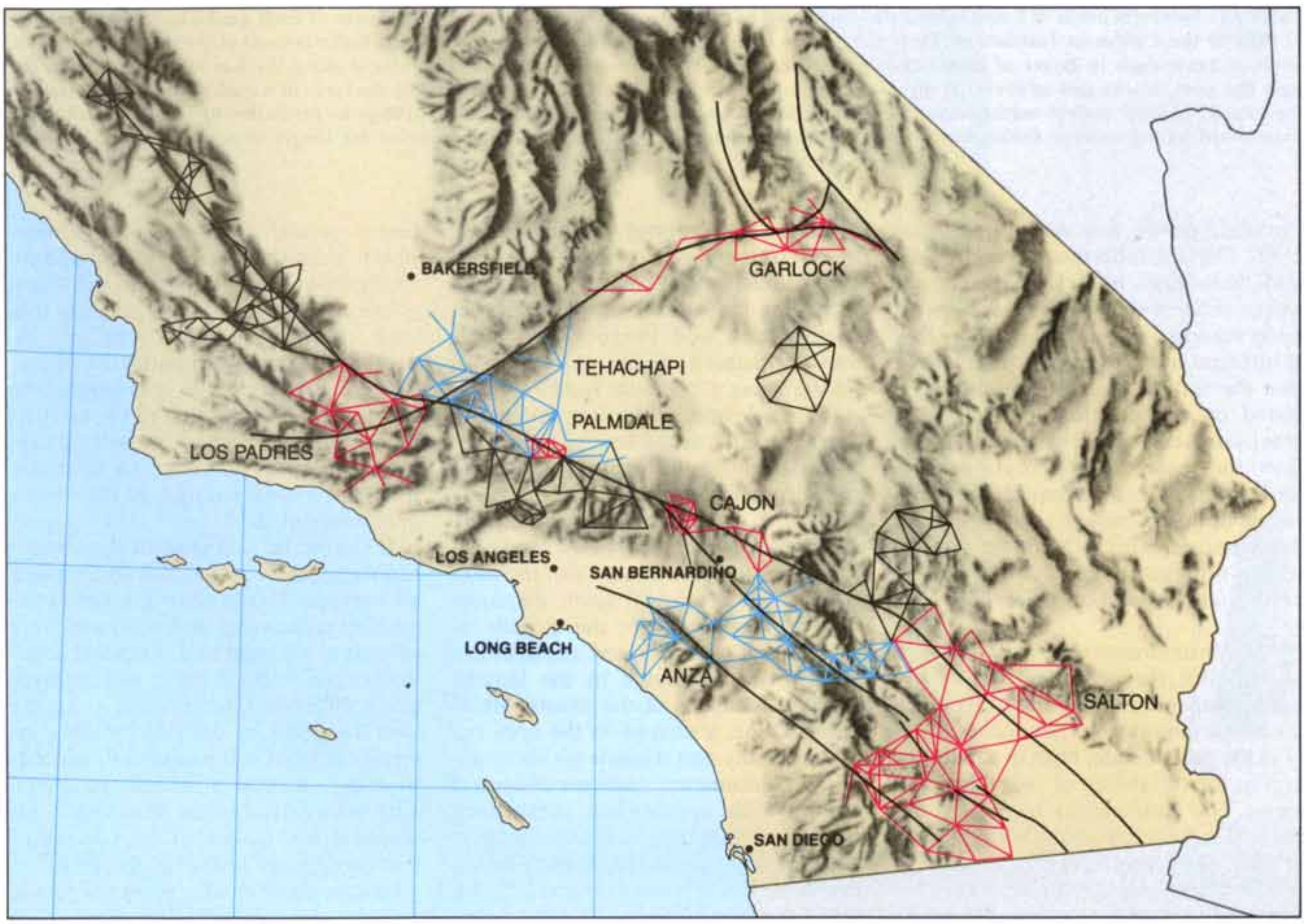
It should be said that all these techniques get their results through repeated surveys. Hence they are time-consuming, expensive and representative of only a regional and temporal average at one point in time. Accordingly great effort has been spent over the past 15 years to develop reliable instruments that will measure strain continuously at one point on the earth. The effort establishes that the crust within a few meters of the surface of the earth is so noisy, in the sense of spurious data, that significant strain signals are masked. The sources of noise include daily and seasonal temperature changes, rainfall, fluctuations in barometric pressure and fluctuations in the water table. Thus early hopes for an inexpensive instrument that could be widely deployed, much like a seismograph, have so far gone unfulfilled.

One instrument capable of continuous measurement of strain is a creep meter. It consists of a wire, say 30 meters long, stretched down the length of a pipe placed a few meters deep in the earth, with a strain gauge at one of its ends. Another instrument offering promise is a borehole strain meter designed by I. Selwyn Sacks and his colleagues at the Carnegie Institution of Washington. In essence it is a fluid-filled balloon fitted tightly into a deep well. The amount of fluid extruded from the balloon in response to crustal compression or expansion is measured continuously by electronic means. Several such instruments are now installed in California. Observations of the water level in wells and of geochemical phenomena such as the concentration of radon gas in well water have proved to be equivocal at best. Geodetic measurements made from satellites are improving; in the future they may offer measurements of the movement of crustal blocks precise enough for employment in earthquake prediction.

What, then, do the available measurements show about the real-time geology of California? How do they aid the effort at short-term prediction? The year 1980 can be taken as a benchmark. Among the geodetic changes observed before 1980 one change was notable: a part of southern California was rising. At the Geological Survey, Robert O. Castle and his colleagues reviewed several decades of data from the U.S. National Geodetic Survey and other sources. They concluded that from 1960 to 1962 the region around Palmdale, about 50 kilometers north of Los Angeles, rose about 25 centimeters. From 1972 to 1974 the uplift increased by a further 15 centimeters. Moreover, the maximum of the uplift shifted 125 kilometers to the east-southeast and the area of uplift expanded southeastward, nearly reaching the Salton Sea. The displacement of that large a volume of crust suggested a great earthquake might be imminent.

Elsewhere in the world—for example at Niigata in Japan—the regional warping of crustal rock had preceded an earthquake. Furthermore, the uplift was occurring in an elliptical region overlying much of the part of the San Andreas fault whose rupture produced the great earthquake of 1857. To be sure, some of the older leveling data (that is, the surveying of land elevation) underlying the discovery of the Palmdale uplift have been questioned. As a result new surveys are done so as to minimize the potential for systematic error. Perhaps the most conclusive evidence that crustal strain is really being detected is the work of Robert Jachens and his colleagues at the Geological Survey. They found a strong correlation among three independent sets of measurements—the local strength of the earth's gravitational field, the changing elevation and the changing strain itself—in southern California over a five-year period.

Other geodetic data suggested that



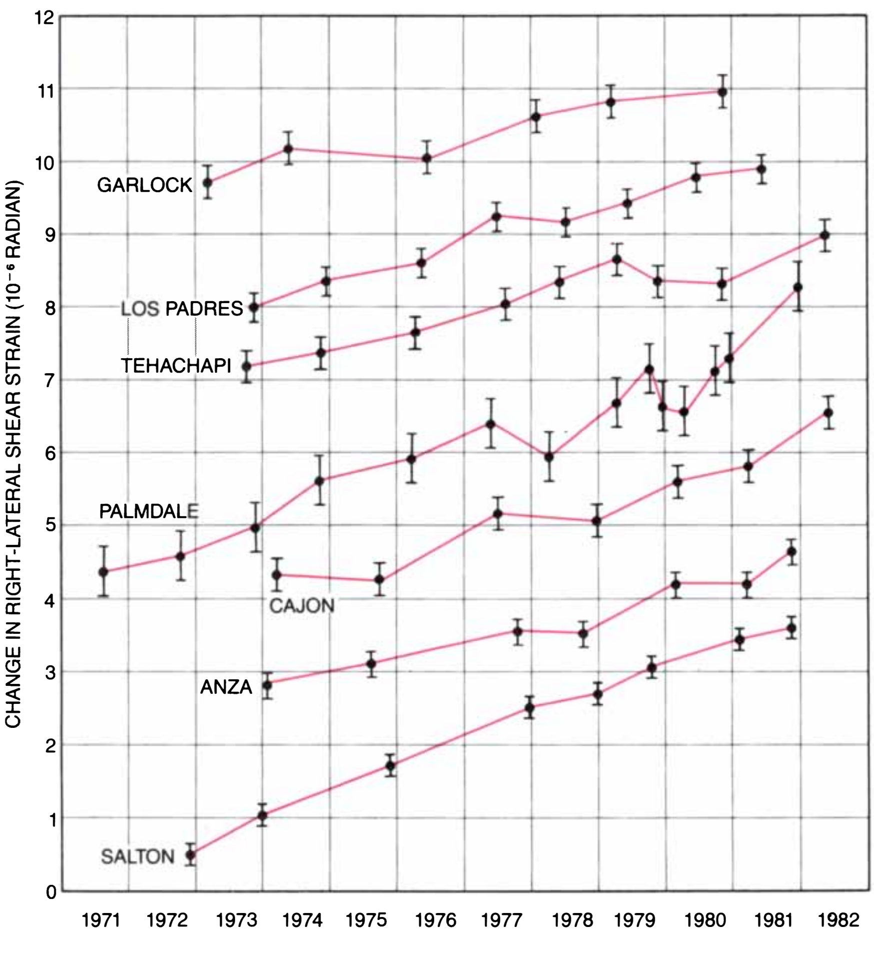
LASER DISTANCE-RANGING NETWORKS form a set of triangles athwart the San Andreas fault in southern California. In each network the time of flight for laser beams yields a measure of distance over baselines as long as 30 kilometers, and so enables investigators to tell from changing distances that strain is accumulating in the crust, perhaps portending an earthquake. Measurements of right-lateral shear strain—the elastic distortion tending to cause displacement on the San Andreas fault—are shown in the chart at the right; the measurements were made by James C. Savage and his colleagues at the Geological Survey. The measurements all tend

further crustal changes were under way—ones that tended to increase the potential for slip on the San Andreas fault. Surveys of horizontal strain made by James C. Savage and his colleagues at the Geological Survey during a 10-year span showed a long-term pattern of crustal shortening north to south. Then, in 1978 or 1979, the strain pattern altered in such a way that it increased the shearing stress along the direction of the San Andreas fault. Clearly slip along the fault was now more possible than before, but the threshold for slip remains unknown.

Also in 1979 measurements of the concentration of radon gas in two wells along the northern margin of the Los Angeles basin showed marked increases. No more than a year of baseline data had been accumulated. Still, the increase was judged to be significant. For one thing, it had been reported from the Soviet Union and the People's Republic of China that such increases preceded several massive earthquakes. Presumably the gas was being released from deforming crustal rock. Taken together, the long-term recurrence data and the observed geophysical changes raised concern in 1980 that a great earthquake might be imminent in southern California.

Since then, however, the effort to formulate a specific prediction for southern California has become rather problematic. Several sites of earthquake potential are recognized. More frequent but smaller earthquakes than the one in 1857 are given an increased likelihood. Certain segments of the San Andreas fault have never ruptured in historical times; the strain accumulated along those stretches may make them the likeliest sites for the next great earthquake. (Some investigators conclude that the part of the southern San Andreas fault southeast of San Bernardino is a very likely site.) Finally, there are large faults in California distinct from the San Andreas fault. They too are capable of generating large earthquakes, and in fact they seem to do so, although at average recurrence intervals longer than 1,000 years. Some of these faults are exposed at the surface; others are buried. A fault some 10 kilometers deep produced the moderate but damaging earthquake at Coalinga in 1983.

A major earthquake-prediction experiment has now been mounted by Allan G. Lindh and his colleagues at the Geological Survey along the Parkfield segment of the San Andreas fault in central California. There earthquakes of magnitude 5.5 to 6 have occurred about once every 22 years, most recently in 1966. (The only exception in more than a century has been the quake in 1934, which came 10 years early.) An array of instruments including creep meters, borehole strain meters and a two-color laser distance-measuring device is now in place. The hope is to learn exactly what geophysical changes and premonitory signals occur before the earthquake expected in 1988, plus or minus a few years. In 1980 William H. Bakun of the Geological Survey predicted—correctly—the location and magnitude of the magnitude-6 earthquake that occurred near Morgan Hill, Calif., in April of last year. No premonitory evidence was detected that would have allowed a short-term prediction of the timing of the quake.



upward, indicating the accumulation of strain. Deviations from a steady increase (accelerations in particular) are conceivably earthquake precursors. Crustal materials cannot withstand a shear strain greater than about 10^{-4} radian, where a radian is an angle of approximately 57 degrees; hence shear strain accumulating in the crust at a rate of 5×10^{-7} radian per year implies an interval for earthquake recurrence on the order of 200 years.

How can the emerging understanding of why—and when—great earthquakes might occur be best used to benefit society? Among the long-term strategies, the incorporation of earthquake resistance into building design and construction is the first line of defense. In California building codes are administered locally but are based on recommendations made by the state's Structural Engineers Association. The recommendations reflect engineering analyses and also experience gained empirically, through observations of how various buildings actually performed during earthquakes. Thus the building codes incorporate a set of minimum standards. In addition state legislation imposes special conditions on schools, hospitals and dams and requires special studies for buildings to be situated near active faults. In general, major construction projects such as tall buildings in Los Angeles and San Francisco and special projects such as nuclear reactors receive particular attention. At the other end of the spectrum most single-family wood-frame buildings are inherently resistant to damage from earthquakes.

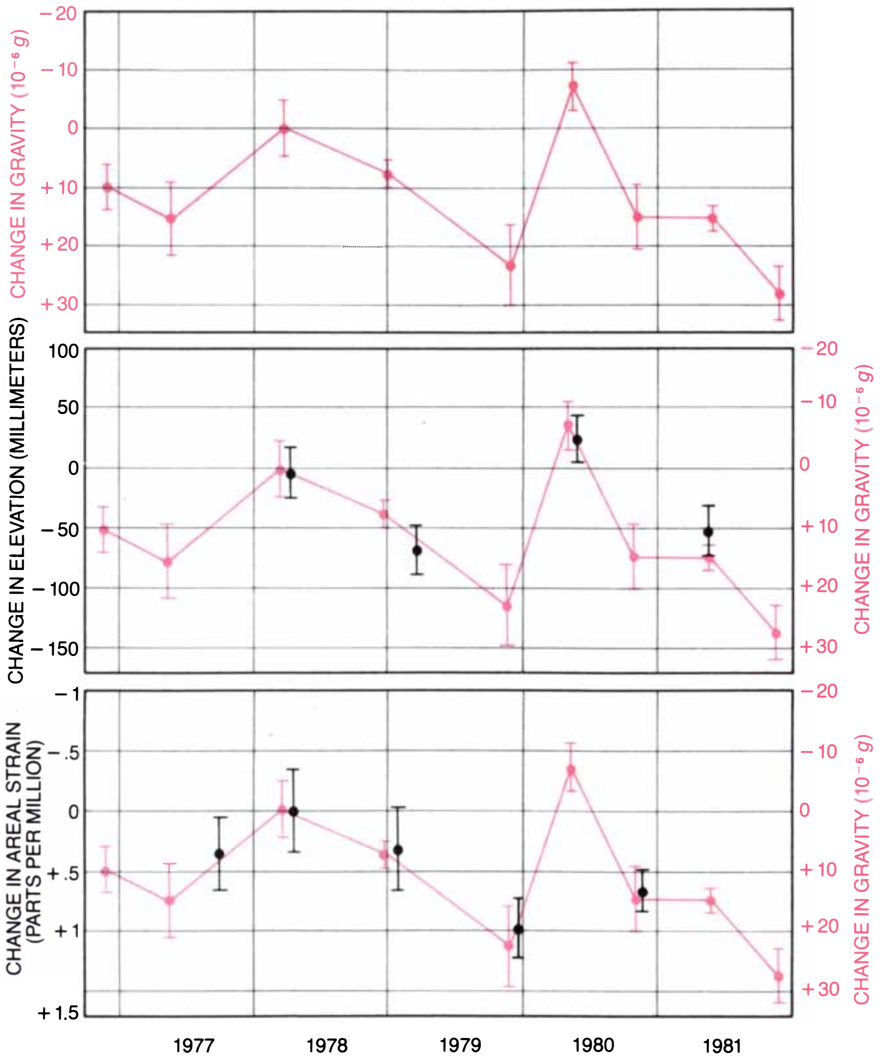
On the other hand, architectural fashions change, and so do the available materials and practices. Hence at any time large numbers of existing buildings fail to conform to the most recent codes or to the state of the art. Moreover, buildings of intermediate size are often built only to minimum code requirements without special provisions for earthquake resistance. Such buildings represent a large proportion of the total inventory of structures in a society. Further still, the care with which an engineering design is implemented—the quality control and the attention to detail, say in the placement of reinforcing steel and the making of welds—is crucial to the success of the design. Sometimes such care is sacrificed in the effort to speed construction or reduce costs.

Many communities in California are implementing plans for the use of land that minimize the potential hazards of earthquakes. Plainly it is imprudent to build structures for human occupancy astride faults or on landslides that are likely to shift catastrophically during a quake. Parks, roads and other open-space uses of such areas pose little threat to life and safety.

Short-term earthquake predictions could increase the margin of safety, particularly for people who live or work in buildings that do not meet the current standards for earthquake resistance or people near a facility whose failure would be potentially hazardous. Some simple strategies invoked by a short-term prediction would include the evacuation of hazardous structures—not entire cities—for a few hours or days, the parking of ambulances and fire trucks outside their garages and the alerting of critical personnel for special duty. At a nuclear power plant additional precautions might range from drills and a review of disaster-response plans to more expensive actions such as the reduction of power or even the temporary shutdown of the plant. Similar opportunities face many industries. The simple removal of fragile inventory—say electronic equipment, chemicals and pharmaceuticals—from shelves could reduce losses dramatically.

As a result of state legislation and state and Federal funding an experimental effort, the Southern California Earthquake Preparedness Project, or SCEPP, was established in 1980. The project is developing plans for using predictions effectively to reduce the social and economic impact of earthquakes, for responding to unpredicted earthquakes and for guiding recovery from an earthquake. Meanwhile administrative mechanisms for issuing information about the likelihood of earthquakes in California are being developed. Federal authority for issuing geologic-hazard information including earthquake predictions rests with the director of the U.S. Geological Survey, who works closely with his state counterpart, the chief of the California Division of Mines and Geology. Both state and Federal panels of earth scientists will assist in judging the scientific validity of a prediction.

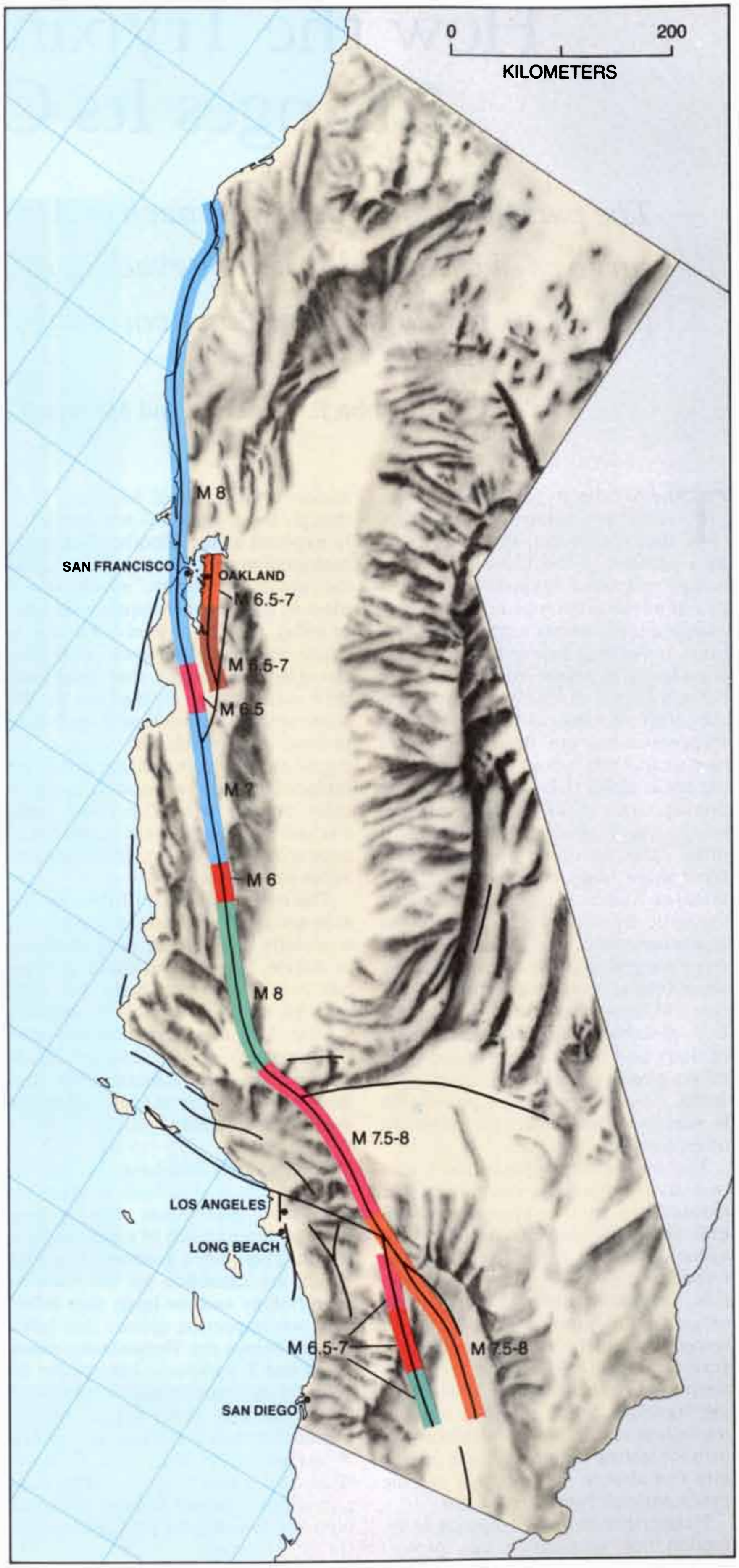
Response to the information (or to an earthquake itself) is the responsibility of the Federal Emergency Management Agency and the State Office of Emergency Services. To date the Geological Survey has issued two official earthquake "watches" (a form of advisory), one for a great earthquake in southern California and one for a moderate but potentially damaging earthquake in the Mammoth Lakes region of eastern California. A series of earthquakes of magnitude 5 to 6 actually followed the latter watch. In several further instances geophysical and geologic anomalies suggestive of a pending earthquake have been reviewed at both state and Federal levels but have not been considered sufficient cause for the issuance of a warning.



GRAVITY, ELEVATION AND HORIZONTAL STRAIN in the crust at Tejon Pass in southern California have been varying together. The charts plot the strength of the local gravitational field (*top chart*), the height of the terrain (*middle chart*) and the strain in the crust, as measured by its changing surface area (*bottom chart*). The correlations among the three suggest that each of these independent measurements reflects a real pattern of crustal fluctuation and not mere "noise" in the local monitoring of the earth. Parameters such as these are being investigated as possible signals foretelling earthquakes. The three sets of measurements were made by Robert Jachens and his colleagues at the Geological Survey.

Earth scientists are gaining confidence in their ability to make long-term predictions of the probability of an earthquake. The next steps will be the efforts to develop integrated systems of measurement and rapid, automated data collection and analysis and apply the analyses to the complexities of crustal deformation. The conversion of long-term predictions into short-term predictions and thence into warnings will pose difficult problems. For example, the long-term evidence suggests that a major earthquake in southern California is imminent. Nevertheless, the short-term variations in geophysical parameters now observed in southern California remain unexplained. Are these variations the precursors of the quake or are they normal fluctuations in the crust? We do not yet have enough experience to say. Many earth scientists lose sleep worrying about how to communicate their concern to the public without raising unnecessary alarm. In the People's Republic of China, where some earthquake predictions have been correct, there have also been many failures. Some of the variations observed today in California might well have been enough to trigger a short-term prediction if they had happened in China.

What will earth scientists do when they are faced simultaneously with three or four different types of anomaly? Such anomalies might include a sequence of small to moderate earthquakes along a previously quiescent segment of the San Andreas fault and the start of fault creep at a time when the accumulation of horizontal strain is accelerating. The investigators may well be reluctant to issue a warning: the costs of a false one are potentially great. Citizens may well be reluctant to respond to a warning: there was reluctance prior to the Mount St. Helens eruptions, some of whose signs were quite visible. Nevertheless, the potential consequences of the next great earthquake in California argue that all reasonable efforts be made to predict the event and that all efforts be made to prepare for the prediction. Such are the birth pangs of real-time geology.



PROBABILITY of an earthquake sometime in the next 30 years along segments of the San Andreas fault was estimated by Allan G. Lindh and his colleagues at the Geological Survey. Each segment has a different probability because its circumstances differ. Some segments have not been yielding to the movement of the Pacific plate past the North American plate; along those segments the danger of the release of strain by a massive earthquake is presumably at its greatest. Other segments have been "creeping," while still others have been releasing strain in modest quakes. All the predictions are for earthquakes associated with the rupture of a given fault segment. Each such quake has a characteristic magnitude (*M*).

How the Trypanosome Changes Its Coat

The parasite, which deprives much of Africa of meat and milk, survives in the bloodstream by evading the immune system. Its trick is to switch on new genes encoding new surface antigens

by John E. Donelson and Mervyn J. Turner

The African trypanosome is a microscopic animal, a protozoan, that spends part of its life cycle as a parasite in the blood of human beings and other mammals. There it causes a fatal neurological disease, trypanosomiasis, whose final stage in humans is sleeping sickness. The disease is endemic in a vast region of Africa defined by the range of the tsetse fly, the intermediate host that carries the trypanosome from one mammalian host to another. About 50 million people are at direct risk of contracting the disease; some 20,000 new cases are reported every year and thousands of other cases no doubt go unreported. Even more important than the direct threat to human beings is the fact that domestic livestock are susceptible to trypanosomiasis. Between them the trypanosome and the tsetse fly make some four million square miles of Africa—an area larger than that of the U.S.—uninhabitable for most breeds of dairy and beef cattle. Having little or no access to meat or dairy products, most of the human population is malnourished and susceptible to other diseases.

The key to the trypanosome's success is its ability to circumvent the mammalian immune system. A mammal ordinarily defends itself against viruses, bacteria or protozoa such as trypanosomes by manufacturing specific antibodies directed against the antigens, or "nonself" molecules, it recognizes on the foreign organism's surface; the antibodies bind to the antigens and neutralize or kill the invading organism. Some antibody-producing cells persist in the bloodstream and provide lasting immunity. Such immunity can also be elicited by a vaccine that mimics a natural infection.

Neither the immune response to infection nor vaccination can protect against trypanosomal infection. Even though these parasites are continually exposed in the bloodstream to the mammalian immune system (unlike the malaria parasite, which spends most of its life cycle sequestered within cells), they have evolved a way to evade the host's defenses: they keep changing the antigen that constitutes their surface coat. By the time the immune system has made new antibodies to bind to new antigens, some of the trypanosomes have shed their coat and replaced it with yet another antigenically distinct one. The host's overworked immune system is unable to cope with the infection, and so the parasites proliferate.

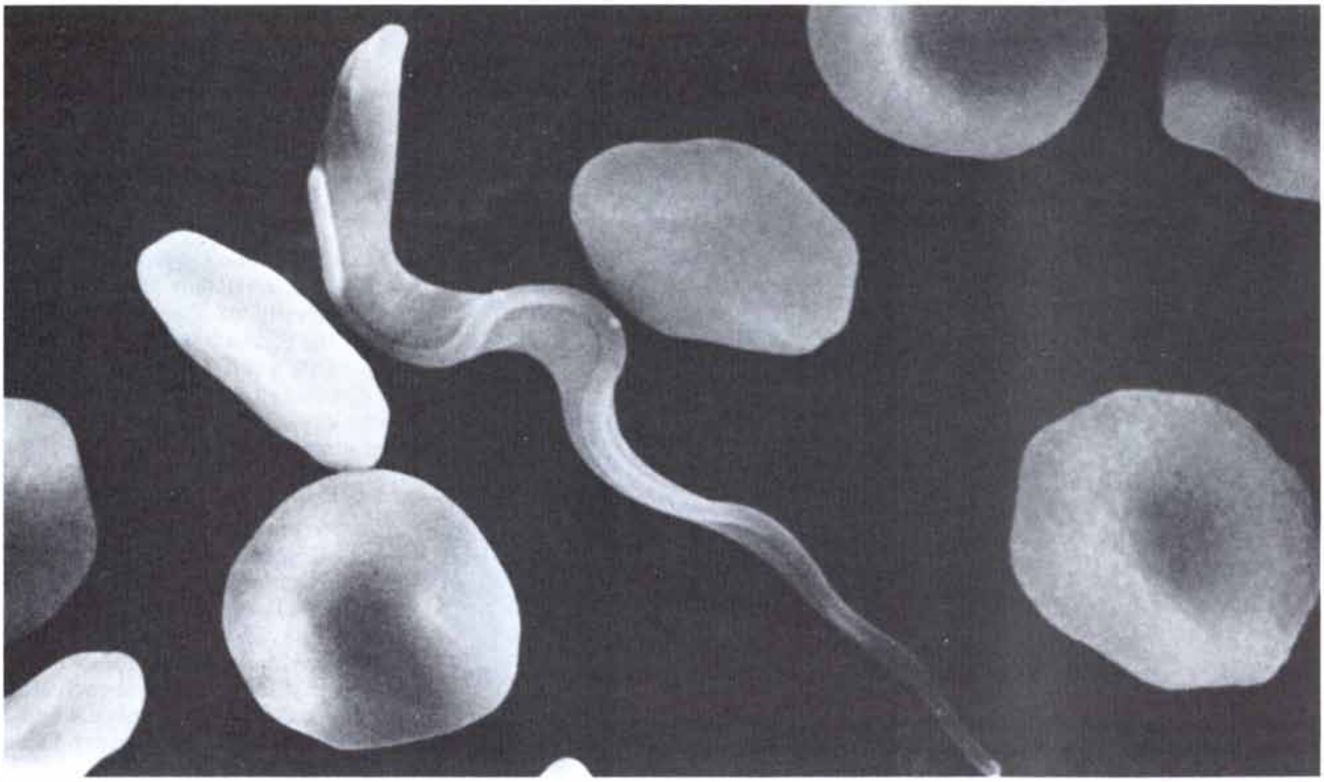
The molecular basis of this remarkable antigenic variation is under intensive study in a number of laboratories in Africa, Europe and the U.S. What has been learned suggests that there may be no way to help the immune system deal with trypanosomes once the parasites are established in the mammalian bloodstream, but that there may be other approaches to prevention or treatment.

The parasites that have evolved this very effective defense strategy are unicellular protozoans ranging from 15 to 30 thousandths of a millimeter in length. There are a number of species, which are classified on the basis of morphology and the hosts they infect. The two important species that infect human beings are *Trypanosoma rhodesiense* and *T. gambiense* (named for the colonial territories where they were first identified). Three species are important for their infection of livestock: *T. congolense, T. vivax* and *T. brucei.* Because it is easy to grow in laboratory animals and cannot survive in human blood, *T. brucei* is the preferred species for investigation.

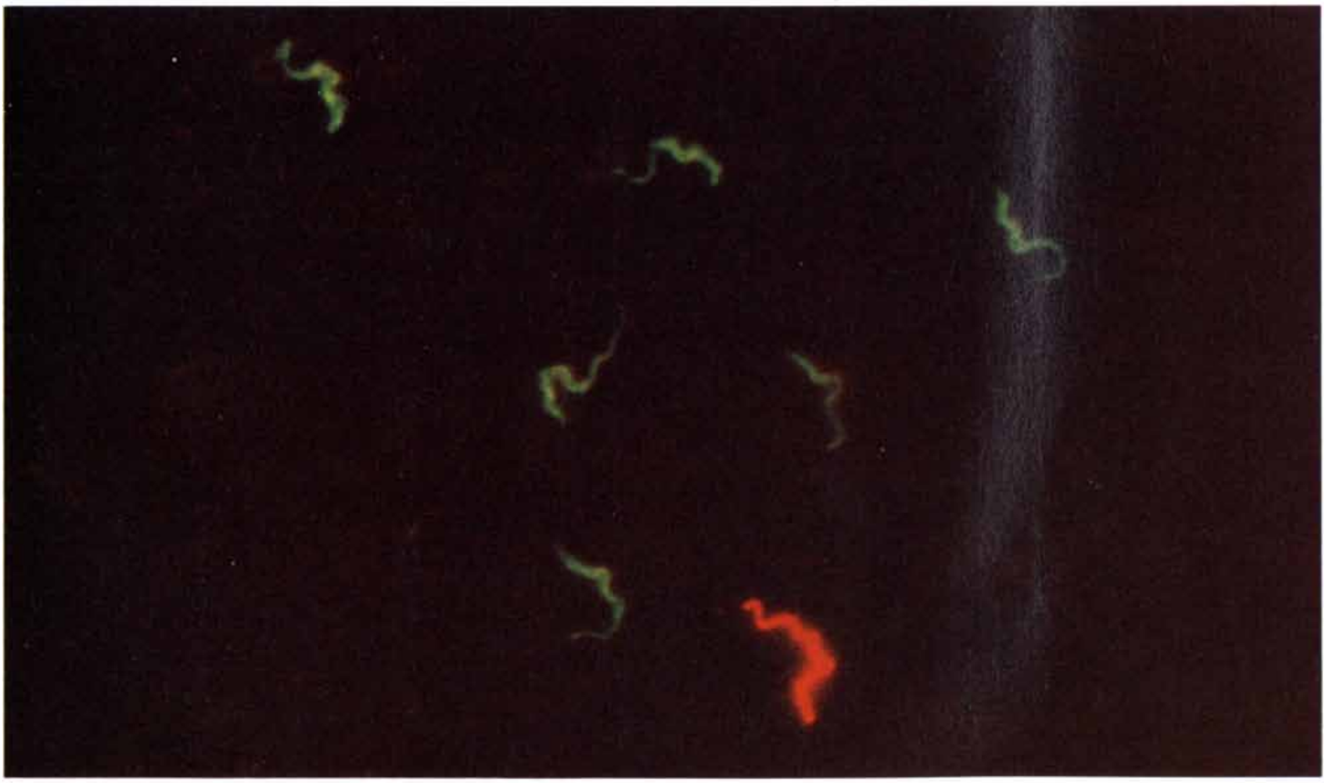
Trypanosomes, like many other parasites, assume different forms at different stages of their complex life cycle. The trypanosomes ingested by a tsetse fly along with an infected mammal's blood lodge in the fly's midgut, where they begin to undergo a series of biochemical and structural changes; in the process they lose their surface coat. After about three weeks they appear in the fly's salivary glands as the metacyclic form, which again carries a surface coat.

When the fly bites a mammal, metacyclic trypanosomes are introduced into the host's bloodstream, where they rapidly differentiate to a form that can proliferate. In humans the resulting disease can be either acute or chronic, depending on the infecting species. In both forms the disease first affects the blood vessels and lymph glands, causing intermittent fever, a rash and swelling. It is at this stage that the continuing battle with the host's immune system begins. Later the parasites invade the central nervous system; inflammation of the outer membranes of the brain leads to lethargy, coma and ultimately death.

The first indication that something keeps changing in the course of a trypanosome infection came early in this century. Physicians had observed a marked periodicity in the temperature of patients with trypanosomiasis. In 1910 the English investigators Ronald Ross and David Thomson, examining blood specimens taken from a patient every few days, noted that the changes in temperature were paralleled by a sharp rise and fall in the number of parasites in the blood. In reporting this finding Ross and Thomson quoted a suggestion, made by an Italian physician named A. Massaglia the year before, that "trypanolytic crises are due to the formation of anti-bodies in the

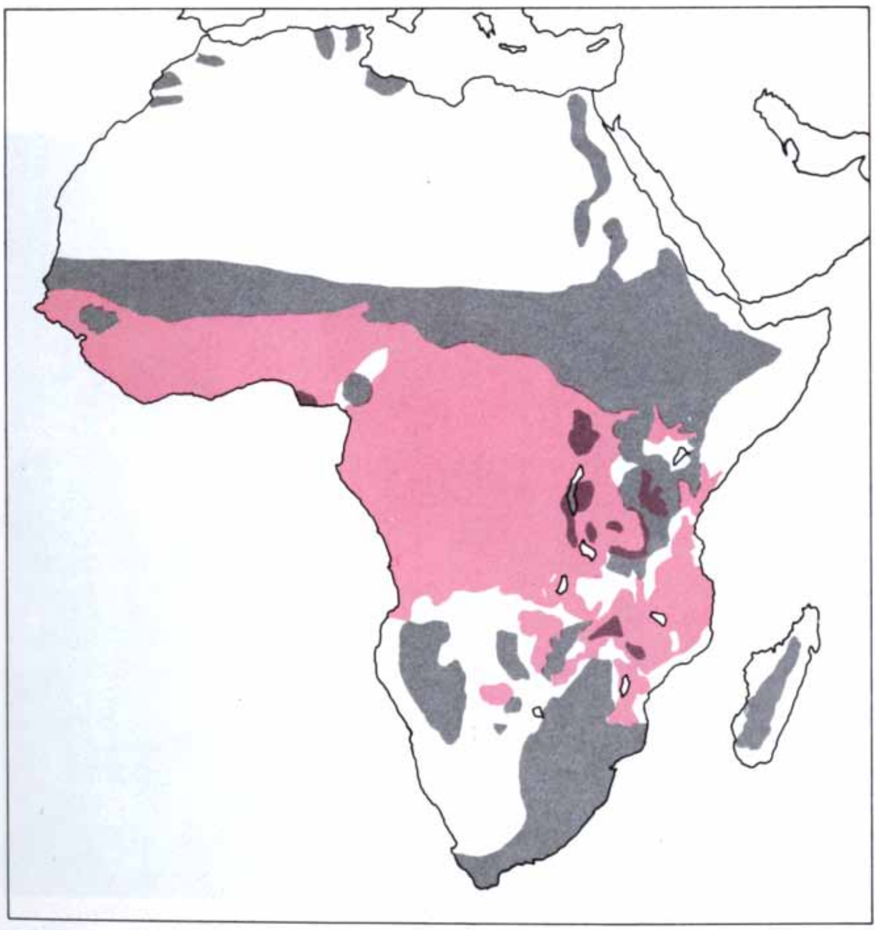


TRYPANOSOME and red blood cells are enlarged 5,500 diameters in this scanning electron micrograph made by Steven T. Brentano of the University of Iowa. The parasite, which is introduced into a mammal's bloodstream by the bite of the tsetse fly, is a unicellular animal (a protozoan) with a single flagellum extending along one side. The parasite's surface is covered with variable surface glycoproteins (VSG's). They are antigenic: the mammal's immune system makes antibodies that bind to them, killing the parasite. A trypanosome can change its surface coat, however, giving rise to a new parasite population that evades the host's antibody defense.



ANTIGENIC VARIATION is made visible in this double-exposure fluorescence micrograph made by Klaus M. Esser of the Walter Reed Army Institute of Research. Most of the trypanosomes are green because the antibody that recognizes and binds to their surface VSG has been labeled with a dye that gives off a green glow under ultraviolet radiation. One parasite is red. It has changed its surface coat and carries a different VSG, which is recognized by a different antibody: one that is linked to a red fluorescent dye.

TSETSE FLY'S proboscis and distended abdomen are red with blood ingested from an experimental animal in this photograph made by Edgar D. Rowton of the Walter Reed Army Institute. The fly is the trypanosome's intermediate host and vector, in which the parasite goes through several developmental stages before it is injected into a mammalian host.



TSETSE FLY AND PARASITE together make some four million square miles of potential grazing land in Africa uninhabitable for most livestock. There is virtually no overlap between the range of the tsetse fly (*color*) and the cattle-raising regions of Africa (*gray*).

blood. A few parasites escape destruction because they become used or habituated to the action of these antibodies. These are the parasites which cause the relapses." It was to be more than 50 years before this perceptive early insight could be confirmed and explained.

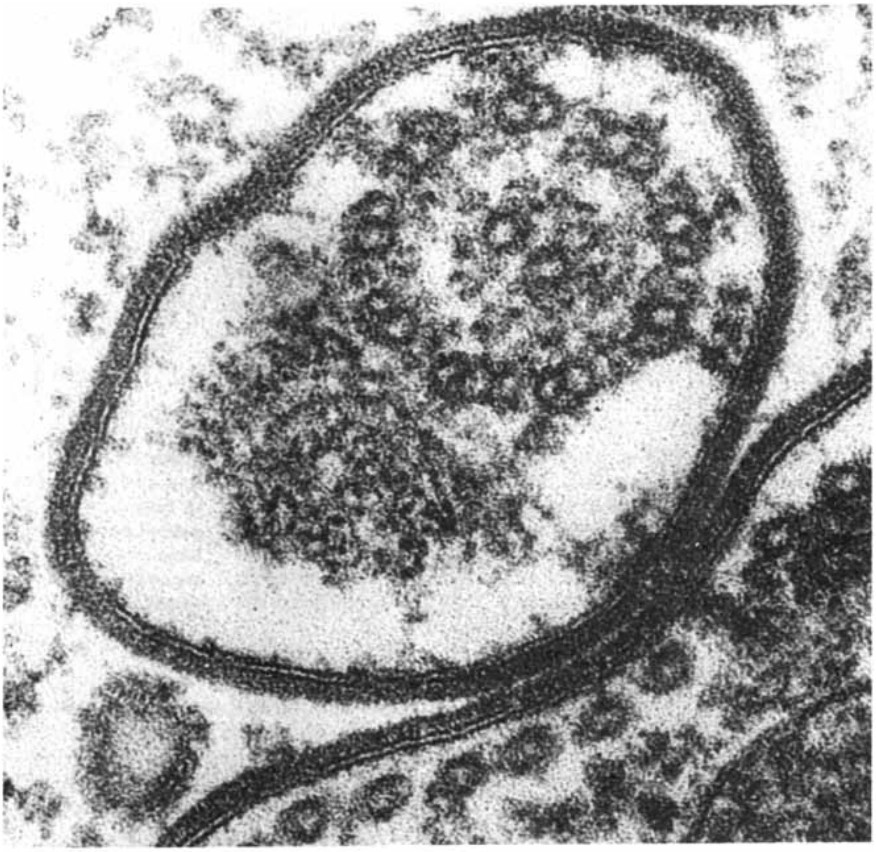
The explanation began to emerge in 1965, when Keith Vickerman of the University of Glasgow first described the thick surface coat covering the parasite's cell membrane. Soon it was discovered that individual trypanosome clones (populations descended from a single ancestral cell) have different surface coats. In 1968 Richard W. F. Le Page of the Medical Research Council's Molteno Institute in Cambridge, England, analyzed the antigenic surface proteins isolated from a number of clones. He found that each clone displayed a biochemically different protein; the differences were so marked that they suggested each antigen must represent the expression of a different gene. (A gene is expressed by a cell when the DNA of the gene is transcribed to make a strand of messenger RNA, which is subsequently translated to make a protein.)

During the mid-1970's George A. M. Cross and his colleagues at the Molteno Institute generated four different clones from a trypanosome population infecting a laboratory animal. They showed that the surface coat consists of a matrix of identical glycoprotein molecules (proteins to which carbohydrate groups are attached) and that the glycoproteins are the same in all individuals in a single clone. Having determined the sequence of the amino acids (the subunits of protein chains) at the beginning of the glycoproteins of their four clones, they noted that the sequence was different in each case; the observation supported Le Page's proposal that different surface antigens must be encoded by different genes. These antigens are now called variable surface glycoproteins, or VSG's.

Early in the course of an infection the immune system generates antibodies shaped to bind to the particular VSG's it "sees" on the surface coat of invading parasites. The antibodies kill perhaps 99 percent of the original parasite population. A few individual trypanosomes escape, however, because they have turned on a different VSG gene and are covered by a new coat of VSG's to which the available antibodies cannot bind. These variant individuals give rise to a new population expressing the new set of VSG's; the new population grows while the immune system raises another set of antibodies, which eventually succeed in again killing some 99 percent of the parasites. By that time, however, a few of the parasites have changed their coat again, and therefore a new population proliferates. So it goes until the host dies. The available evidence suggests that the switch from one VSG to another takes place spontaneously. The host's immune system does not induce the switch but rather selects (by its inability to deal quickly with a particular new antigen) a variant that initiates a new population.

The total potential VSG repertory of a trypanosome is not known. In controlled experiments trypanosomes derived from a single parent cell have generated more than 100 distinct VSG's, giving no indication that their complement of VSG genes has been exhausted. It has recently been estimated that a single organism has at least a few hundred and perhaps as many as 1,000 or so VSG genes. (This means that from 5 to 10 percent of the parasite's total genetic capacity is devoted to antigenic variation.) In the wild the combined gene pool of all trypanosomes probably supplies genetic information to generate a virtually infinite number of antigenically distinct VSG's.

What is the structure of the VSG and how is it attached to the cell membrane? How does the parasite manage to express one VSG gene at a time (and only one) out of a repertory of hundreds of such genes? In order to learn something about the VSG's and about how they are displayed sequentially by the trypanosome, a number of research groups have been exploiting procedures based on recombinant-DNA technology.



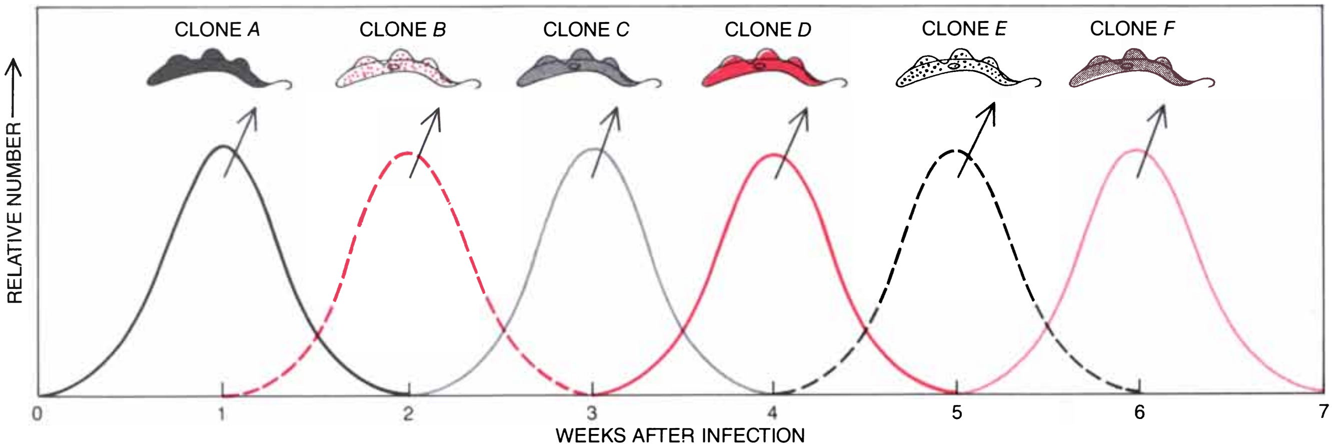
TRYPANOSOME'S SURFACE COAT of VSG's is visible as a diffuse, dark layer in an electron micrograph made by Laurence Tetley and Keith Vickerman of the University of Glasgow. A cross section of the parasite's body and flagellum has been enlarged about 190,000 diameters. The double membrane just inside the surface coat is the cell membrane.

The first step is to isolate the messenger RNA (mRNA) from a trypanosome clone. The mRNA is copied to make what is called complementary DNA (cDNA), which is combined with a carrier DNA and introduced into bacteria. By manipulating the bacteria harboring the recombinant DNA, it is possible eventually to isolate a cDNA molecule that is in effect a copy of the VSG gene the trypanosome clone is expressing. By determining the sequence of nucleotides (the components of DNA) in the cDNA and translating it according to the genetic code, it is possible to predict the amino acid sequence of the VSG the gene encodes.

Partial or complete nucleotide sequences have been determined for some 15 cDNA's, and analysis of the deduced amino acid sequences reveals that each newly synthesized VSG is made up of about 500 amino acids. The first 20 or 30 of them constitute a signal peptide (a short protein chain) whose function is to help move the nascent VSG across the trypanosome's cell membrane; comparison of the cDNA-predicted amino acid sequence with the actual sequence of a few VSG's has shown that in the process the signal peptide is cleaved off. The sequence of the next 360 amino acids is quite different in most VSG's, and this variable region is presumably responsible for the parasite's antigenic diversity. The last 120 amino acids (at what is called the *C*-terminal end of the protein) are quite similar in various VSG's; on the basis of the degree of similarity in these homology regions, VSG's can be classified into two homology groups.

Ordinarily it is peptide sequences in the *C*-terminal region of a surface protein that anchor the protein in the cell membrane, but the VSG is different. The last 20-odd amino acids of the region are clipped off in the mature VSG and are replaced by a structure containing an unusual oligosaccharide (a complex sugar molecule) that has a role in the anchoring. It seems to be very similar in all VSG's, regardless of their variable-region sequence, because antibodies raised against one oligosaccharide bind to all VSG molecules. Why does this cross-reacting determinant, as it is called, not induce natural immunity and why can it not serve as the basis of a vaccine? The reason is essentially that VSG's are packed into the surface coat in such a way that the immune system is confronted by the variable domain of the protein; the homology domain, including the cross-reacting determinant, is not exposed.

The cross-reacting determinant appears to be part of a larger oligosaccharide molecule, which is linked in turn to a phosphoglyceride carrying two fatty acid chains. It is the fatty acids that penetrate the cell membrane and hold the VSG in place. Why would the trypanosome substitute this complicated structure for the usual *C*-terminal cell-membrane anchor? The reason may be that the ability to release its surface coat rapidly is of central importance to the parasite. The trypanosome has an enzyme that can cleave the link to the fatty acids, thereby releasing the VSG from the membrane. Having the same specialized anchoring molecule, subject to cleavage by the same enzyme, on all VSG's re-



SUCCESSIVE WAVES of parasite proliferation in the blood are characteristic of trypanosomiasis. They result from antigenic variation. A population of parasites, some of which carry a particular antigen, VSG *A,* on their surface, proliferates in the bloodstream for a few days. The immune system raises antibodies against the population's antigens, killing most of the parasites. A few individual parasites survive by expressing new VSG genes that direct the synthesis of new antigens, such as VSG *B;* these parasites give rise to a new population that grows until the immune system manages to raise new antibodies against the new antigens. The cycle is repeated many times in the course of a chronic infection as parasites keep expressing new genes and displaying new VSG's. From each successive population it is possible to isolate individual trypanosomes and from them to grow clones expressing particular VSG's.

gardless of their exact sequence provides a rapid and universal mechanism for stripping off one coat and substituting another.

Although comparison of cDNA sequences has revealed extensive differences among VSG's, it has become clear that a very few amino acid changes can suffice to generate antigenically distinct VSG's. Presumably these changes take place at particular antigenic sites within the 360-amino-acid variable region. To find these sites it will be necessary to know the three-dimensional structure of the variable region in detail. Some progress toward that goal has come recently from X-ray-crystallographic studies done by Don C. Wiley, Douglas M. Freymann and Peter Metcalf of Harvard University, in collaboration with one of us (Turner) at the Molteno Institute.

Five VSG's have been crystallized to date, but most progress has been made in determining the structure of the variable domain prepared from one of them. The resolution so far attained is enough to reveal the parts of a protein chain that are folded into the cylinder-like conformation known as an alpha helix. About half of the variable-region turns out to be in that form.

The variable region crystallizes as a dimer (a double molecule made up of two monomers), and indeed VSG's seem to aggregate as dimers in the surface coat of a living trypanosome as well. The dimer (or at least the resolved half of it) is shown by crystallography to be a bundle of alpha helixes. The core of the bundle is made up of two hairpin-shaped structures, one from each of the component monomers. At one end of the dimer the core interacts with two more helixes to form a six-helix bundle; at the other end the monomers diverge somewhat to form a distinctive head. This highly symmetrical structure must form the framework for the remaining half of the variable-domain sequence, unrevealed so far because it is not an alpha helix. We still do not know how the framework structure is oriented in the membrane (that is, which end is "out"), and so we cannot begin to guess where on the structure the major antigenic sites will be found.

Just what goes on when the trypanosome turns on one VSG gene after another, always synthesizing only one antigenically distinct surface glycoprotein at a time? Experiments designed to answer that question depend again on cDNA's of the kind described above. Now each such cDNA, which is in effect an artificial VSG gene, serves as a probe with which to locate copies of the same gene wherever they may be in the trypanosome genome (the total complement of genetic material).

The total genomic DNA is digested with a restriction enzyme, which cleaves the DNA at a specific site within a particular sequence of nucleotides. The result is that the genomic DNA is broken down into a large number of small fragments, each one slightly different in size. Having been separated according to their size by a process called gel electrophoresis, the fragments are transferred to nitrocellulose paper, to which they bind tightly. A cDNA representing a VSG gene, labeled with a radioactive isotope, is applied to the paper. The cDNA hybridizes with (binds to) any similar sequences it finds among the restriction fragments on the paper. Unbound cDNA is washed away and autoradiography reveals the sites of hybridization. One can thus determine how many different copies there are in a given trypanosome clone of the gene represented by the cDNA probe and whether or not the gene is in a different part of the genome in different clones.

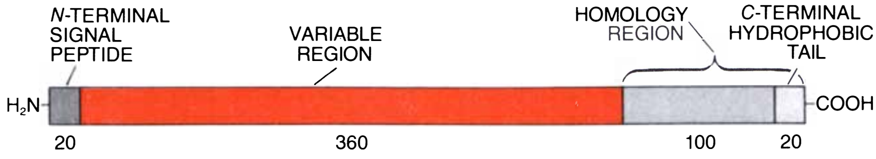
In this manner Piet Borst and his colleagues at the University of Amsterdam and Cross and his associates at the Wellcome Research Laboratories in England demonstrated that when some VSG genes are expressed, an extra copy of the gene is present in the genome. They called it an expression-linked copy. Étienne Pays and Maurice Steinert of the Free University of Brussels went on to show that the mRNA for the expressed VSG gene is actually transcribed from this copy rather than from the original (basic-copy) gene that gave rise to it. Further analysis revealed that an expression-linked copy is always at a particular kind of site: near a telomere (the end of a chromosome). In other words, a single VSG gene out of the total repertory of such genes is expressed when it is duplicated and translocated to an expression site near a telomere; the switch from one VSG to another is often effected by the removal and degradation of one such copy and its replacement by the copy of another gene.

This copy-and-translocate mechanism is not the only source of antigenic variation. John R. Young, Phelix A. O. Majiwa and Richard O. Williams of the International Laboratory for Research on Animal Diseases in Nairobi found another mechanism. They discovered that in some cases the number of different fragments to which a particular cDNA probe hybridizes does not change when the VSG gene represented by the probe is expressed: there is no expression-linked copy. The probe hybridizes instead to a fragment that is different in size in each trypanosome clone, whether or not the clone is expressing the gene.

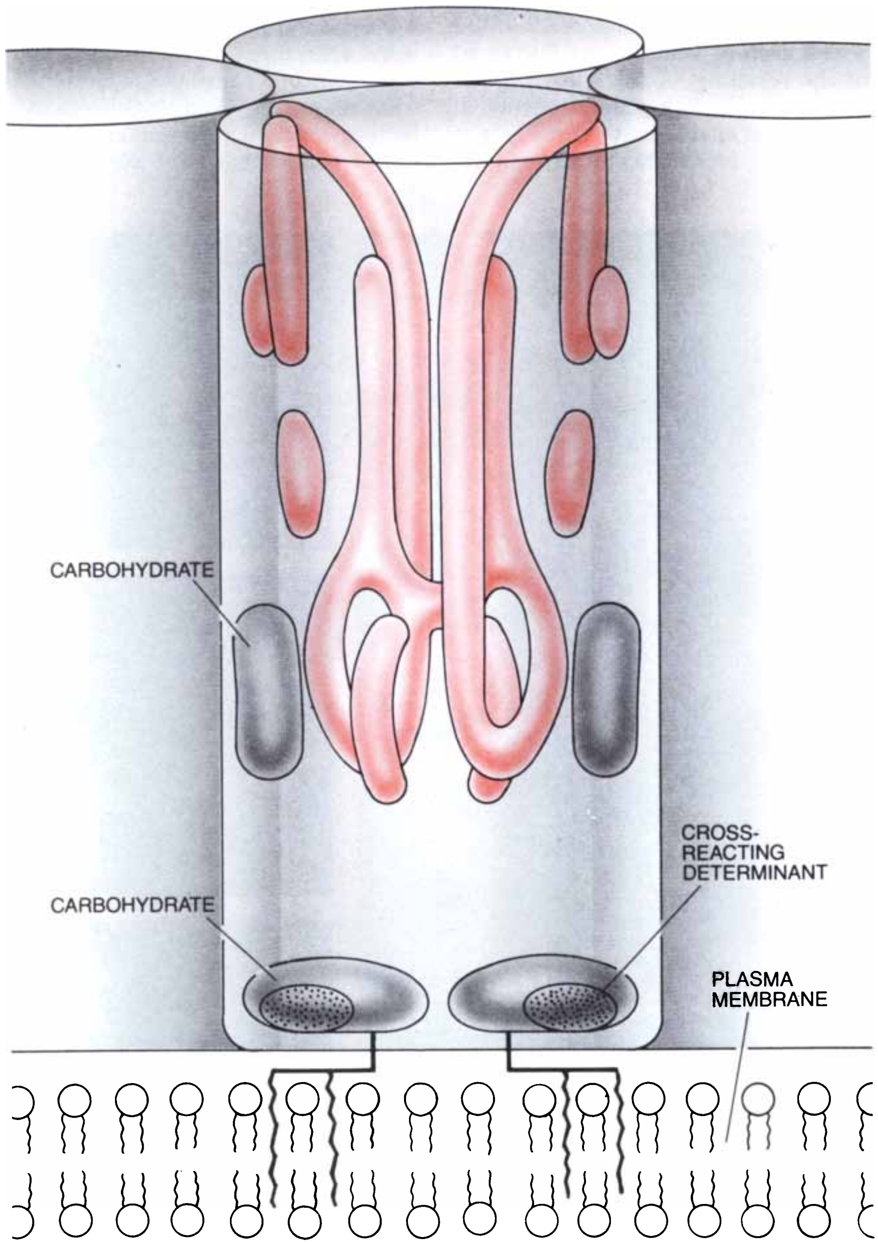
What this means is that some VSG genes are expressed without being duplicated and translocated. These genes turn out to be already at a site near a telomere. Their proximity to a telomere explains why their fragments vary in size. In trypanosomes and in some other organisms it is often the case that a short DNA sequence is repeated hundreds of times near a telomere. The number of these "tandem repeats" between a VSG gene and a telomere varies from clone to clone. As a result the fragment carrying a given telomere-linked VSG gene can be a different size in different clones.

About half of the VSG genes studied to date seem to be telomere-linked ones. Since there are hundreds of different VSG genes in the trypanosome genome, this suggests that there must be at least several hundred telomere-linked VSG genes in the genome, and so there must be hundreds of chromosomes in the parasite's nucleus! (The human chromosome complement is only 46.) Because the parasite has a normal amount of DNA for a unicellular animal, one would expect that some of the parasite's chromosomes must be very small, and indeed some of them are. A student of Borst's, Lex H. T. van der Ploeg, has applied a technique developed by David C. Schwartz and Charles R. Cantor of Columbia University to resolve trypanosomal nuclear DNA into four size classes: minichromosomes about 100,000 nucleotides long, other small chromosomes from two to seven times as long, middle-sized molecules (about two million nucleotides long) and others that are too long to be measured by this method.

Van der Ploeg found VSG genes on chromosomes in every size class. In one case he found a basic-copy gene on a large chromosome and its corresponding expression-linked copy on a middle-sized one, indicating that translocation of a duplicated molecule can take place between chromosomes. In another case a telomere-linked gene was not duplicated but was expressed at its normal site near the end of a large chromosome. These observations supported earlier evidence, accumulated in several laboratories, that there must be more than one site in the genome at which VSG genes can be activated; they are not all translocated to the



PROTEIN CHAIN of a typical VSG is composed of about 500 amino acids. The first 20 or so of these, at what is called the *N* terminal of the protein, constitute a signal peptide, which is cleaved from the protein before the VSG is implanted in the cell membrane. The next 360 amino acids (*color*) constitute the variable region, which is different in each antigenically distinct VSG. The final 120 amino acids at the *C* terminal are quite similar in each of two "homology groups" of VSG's. The last 20 amino acids of this region are cleaved from the chain and replaced by a large molecule that anchors the VSG in the cell membrane.



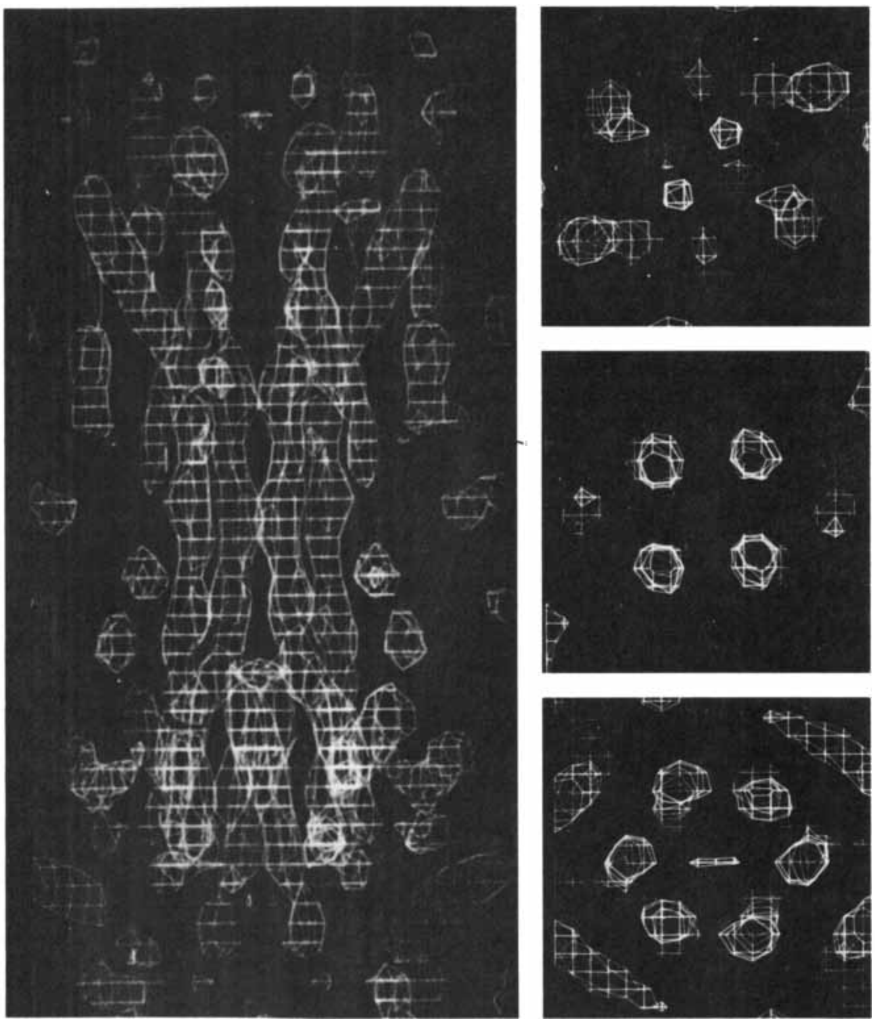
VSG'S OF SURFACE COAT may be assembled as indicated in this somewhat speculative drawing. The structure of part of the variable region of one VSG (*color*) is based on X-ray-crystallographic data (*see illustration on next page*); the full extent of the VSG and the location of its neighbors are indicated by the gray cylinders. The variable region is a dimer, or double molecule, that appears to be a bundle of the protein structures called alpha helixes. Carbohydrate molecules flank the bundle. Two more carbohydrates at the base of the VSG may incorporate a small sugar molecule: the cross-reacting determinant. Fatty acids extending from these two carbohydrates appear to anchor the VSG in the membrane.

same unique expression site near a particular telomere. Although proximity to a telomere is necessary for expression, it therefore cannot be sufficient. Other factors must come into play that select and activate a single VSG gene in a given organism for transcription into mRNA at a given time.

Some of the events attending this activation may contribute to increasing the diversity of VSG's. Pays and Steinert and their colleagues have reported cases in which a functional expression-linked copy was generated not by the duplication of a basic-copy gene but by the "recombination" of segments of at least two different telomere-linked genes, each of which codes for a part of the resulting VSG. If this kind of recombination is a general phenomenon, it must enable the trypanosome to make even more VSG's than its hundreds of genomic VSG genes can specify. It also suggests a reason for the telomeric location of expression sites: highly repetitive stretches of DNA such as the short tandem repeats near telomeres are particularly likely to undergo recombination.

Occasionally an expression-linked copy is not destroyed during the switch to the expression of another gene but instead "lingers" for a while in its expression site. At the University of Iowa one of us (Donelson) has recently shown that the sequences at the boundaries of one particular expression-linked copy are virtually identical with sequences bounding a telomere-linked gene. This, together with similar findings from other groups, suggests that some telomere-linked VSG genes may be previously expressed expression-linked copies that have survived the switch to a new gene. Perhaps two chromosomes exchange regions adjacent to their telomeres, so that an expression-linked copy is removed from its original expression site and placed near a different telomere, where it is available for future expression. Perhaps, on the other hand, a single small segment of DNA in the genome serves as a mobile control element: an enhancer of transcription that can move from one telomere to another and cause different VSG genes to be expressed. In that case an expression-linked copy gene could remain in its expression site but be turned off by the control element's departure.

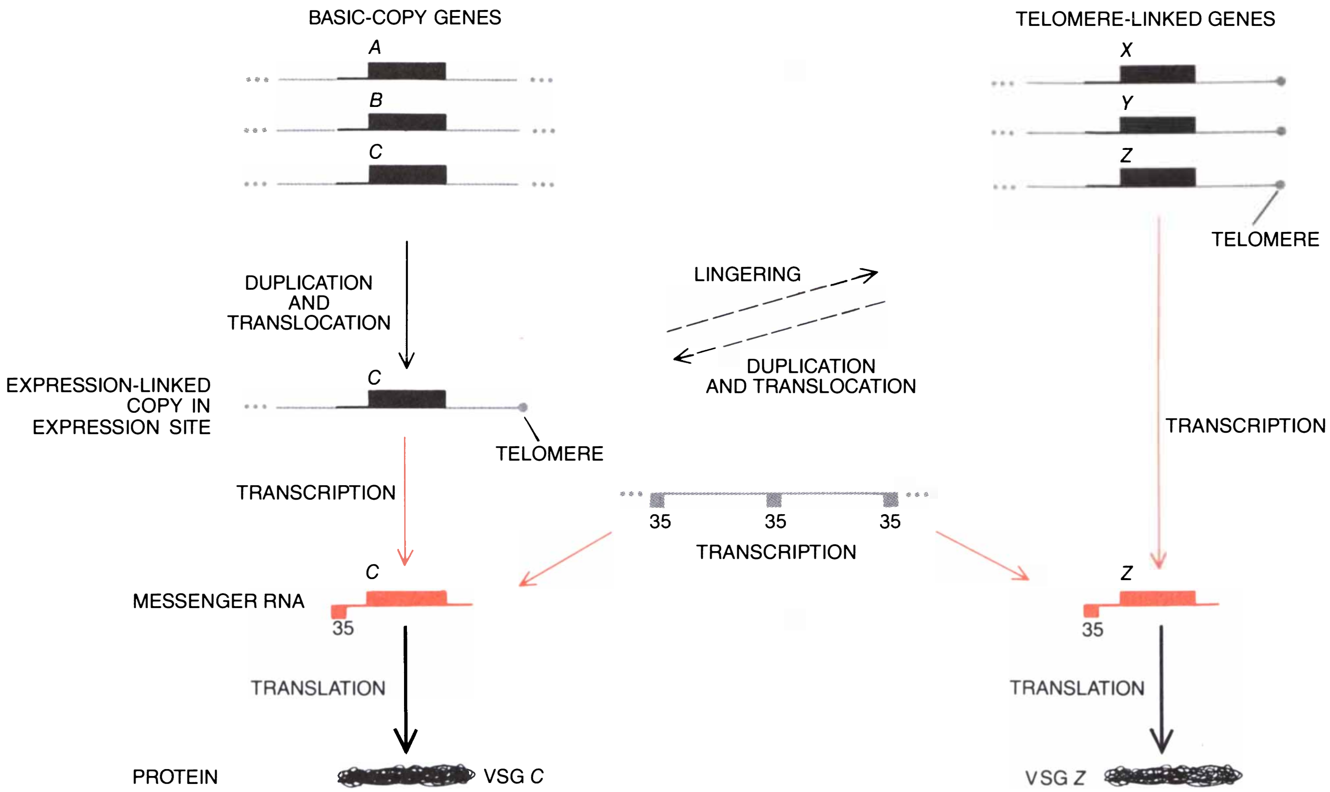
There is another peculiarity of trypanosomes and related organisms. The mRNA's for VSG's (and for many, if not all, other proteins) always begin with the same specific sequence of 35 nucleotides. This sequence is not found in the corresponding gene or anywhere in the DNA surrounding the gene. Instead it is coded for by a repetitive DNA sequence separated from the gene. Small RNA molecules transcribed from this repeated DNA somehow provide copies of the 35-nucleotide sequence for the beginning of each mRNA. The 35-nucleotide sequence presumably has a role in the expression of trypanosome genes, but that role is not yet known.



STRUCTURE of about half of a VSG's variable region was determined by X-ray crystallography by Don C. Wiley, Douglas M. Freymann and Peter Metcalf of Harvard University in collaboration with one of the authors (Turner). The region crystallizes as a dimer. An electron-density map (*left*), a view perpendicular to the axis of the dimer, reveals a bundle of rodlike alpha helixes. There are six helixes at one end (shown here as the bottom, although the actual orientation of the dimer is not known) and from four to six in the middle; the two monomers diverge at the top to form a two-headed structure. The three cross-sectional maps (*right*) are views along the dimer axis near the top, in the middle and near the bottom.

Let us now summarize what is known about the mechanisms of antigenic variation in the trypanosome. There are hundreds of genes encoding VSG's in each organism's genome. They may be in the interior of chromosomes or near telomeres. Only one VSG is transcribed at a time, and that gene is always near a telomere. In order to be transcribed, an interior (basic-copy) gene needs to be duplicated and translocated (as an expression-linked copy) to one of many expression sites, all of which are near telomeres. A telomere-linked VSG gene, on the other hand, need not be duplicated in order to be expressed (although it may in fact sometimes be duplicated). Antigenic diversity may be further increased by recombination. The precise molecular mechanisms triggering the switch from one VSG gene to another are still not known. It would appear, however, that the mechanisms are so complex and varied that it is almost impossible to circumvent them, and so the development of a vaccine against trypanosomes in the bloodstream is unlikely.

It may, however, be possible to develop a vaccine against metacyclic trypanosomes. As we mentioned above, the metacyclic stage is the final developmental stage in the tsetse fly's salivary glands, and it is metacyclic para-



BASIC-COPY GENES in the interior of a trypanosome's chromosomes (*left*) are not ordinarily transcribed into messenger RNA (mRNA). For one of them (*C*) to be expressed it must be duplicated to provide an expression-linked copy. The DNA of the copy is translocated to an expression site near a telomere (the end of a chromosome) and is transcribed into mRNA, which is translated to make the VSG protein. Other VSG genes are already near a telomere (*right*). These telomere-linked genes can be expressed without forming an expression-linked copy, but occasionally they too are duplicated and translocated. An expression-linked copy is usually lost during an antigenic switch but sometimes "lingers" to become an unexpressed telomere-linked gene. Strangely, a 35-nucleotide sequence at one end of the mRNA is not encoded by the VSG gene but is specified instead by a separate region of repeated DNA.

sites that are injected into the bloodstream when a fly bites a mammal. Steven L. Hajduk, J. David Barry and Vickerman at Glasgow and Klaus M. Esser of the Walter Reed Army Institute of Research in Washington found that the metacyclic parasites can display on their surface only a reduced subset of VSG's—perhaps as few as 15. Collaborating with Esser, one of us (Donelson) has studied the cDNA's of several metacyclic VSG's and shown that the proteins have about the same *C*-terminal homology region as bloodstream VSG's; apparently they attach to the cell membrane in the same way. Moreover, the genes for the metacyclic VSG's seem to be telomere-linked ones, like many of the bloodstream VSG genes. It is therefore not yet clear why metacyclic parasites cannot express the full range of VSG's. There may nonetheless be a way to take advantage of their limited repertory or of their switching mechanism in order to make an effective vaccine.

The drugs currently available for treating trypanosomiasis are extremely toxic and also cannot prevent reinfection, but new forms of chemotherapy may yet be found. The trypanosome cannot survive in a mammalian host without its surface coat; a drug that interferes with the phosphoglyceride anchoring the VSG in the cell membrane, or that activates the enzyme releasing the VSG, might therefore be an effective therapeutic agent. Mammalian messenger RNA's do not have the unusual 35-nucleotide sequence described above; a drug that interferes with its synthesis might therefore selectively disable the parasite. Drugs might also be found that act against two subcellular organelles that seem to be unique to trypanosomes. One is the glycosome, a membrane-bounded aggregation of enzymes; the other is the kinetoplast, an appendage of the parasite's single large mitochondrion. A drug interfering with a metabolic function of these organelles, or with some other unique metabolic pathway as yet undiscovered, might kill the trypanosome without harming its mammalian host.

There are some other possible approaches. The tsetse fly can be eradicated briefly in small areas by spraying with insecticides or by the distribution of sterile male flies, but such methods cannot be effective for a region that extends for some four million square miles and more than a score of countries. A few breeds of cattle herded by nomadic tribes do seem to have developed partial resistance to trypanosomiasis. They do not yield much meat or milk, but it may be possible to cross them with more productive breeds. It is also conceivable that wild animals such as the eland or the oryx, which seem not to be harmed by trypanosomes, might be domesticated and take the place of cattle.

The past few years have made it clear that African trypanosomes are superb experimental organisms. Continued investigation of their variable surface coat will provide basic information about such diverse subjects as the control of gene expression, the attachment and functioning of membrane proteins, the structure and replication of chromosomal telomeres and the molecular mechanisms that generate biological diversity.

Fundamental knowledge of this kind should in turn contribute to a truly pressing public health task: the control of trypanosomiasis. The next few years should show if new basic information about trypanosomes can be applied to control and perhaps eventually eradicate trypanosomiasis.

Gamma-Ray Bursters

Intense flashes of high-energy radiation appear unpredictably in the sky. Limited observational data have frustrated attempts to determine their cause, but likely mechanisms have been proposed

by Bradley E. Schaefer

Approximately once per day a burst of very intense gamma radiation emanates from some completely unpredictable part of the sky. The duration of the burst is typically between one second and 10 seconds, although some bursts have been as short as .01 second and others have been as long as 80 seconds. During this time the burst brightens into visibility, varies randomly in intensity and then fades back to invisibility. With few exceptions, no more than one burst has come from exactly the same direction and none has been positively identified with a previously known object.

At their peak these bursts are by far the brightest emissions of gamma radiation in the sky. In fact, if we make reasonable assumptions concerning the distance to the bursts' sources (which are called gamma-ray bursters), it seems that they generate more power per unit of volume than any other object in the known universe. One estimate is that they generate as much energy in one second as the sun does in one week.

In spite of the tremendous amount of energy involved in each emission, however, little is known about gamma-ray bursters. The bursts are unpredictable, and they fall in a region of the spectrum that is extremely difficult to measure precisely. The paucity of observational data allows a great number of proposed models. To date nearly 40 models of bursters have been published, including such exotic ideas as exploding black holes, ultrarelativistic dust grains and the fission of superheavy elements. It is embarrassing that most of these models are consistent with current observational data.

Even with the limited amount of available information, however, astronomers have recently begun to come to a consensus on what the general characteristics of a gamma-ray burster must be. In addition instruments are being built that may answer many of the outstanding questions about the properties of bursts. It may be possible in the near future to determine precisely what causes a gamma-ray burst, thus ending the speculation these mysterious phenomena have aroused since their discovery more than a decade ago.

Gamma-ray bursters were first observed by a set of Vela satellites operated by the U.S. Department of Defense. The satellites, which were designed to monitor Soviet compliance with a treaty banning nuclear tests in space, carried detectors to record the sudden flash of gamma radiation that accompanies a nuclear explosion. After a number of flashes had been detected it was realized that they represented not atomic tests but an entirely new class of previously unknown astronomical objects. The discovery of gamma-ray bursters was soon confirmed by Thomas L. Cline and Upendra D. Desai of the National Aeronautics and Space Administration's Goddard Space Flight Center. Soon thereafter a number of other groups were able to find gamma-ray bursts in data from earlier experiments designed for other purposes.

In the late 1970's a second generation of experiments was begun [see "Cosmic Gamma-Ray Bursts," by Ian B. Strong and Ray W. Klebesadel; SCIENTIFIC AMERICAN, October, 1976]. In order to pinpoint the direction of a gamma-ray burst as accurately as possible, investigators placed a network consisting of roughly a dozen detectors on various spacecraft that were to travel within the inner solar system.

In one of these experiments, called the Konus experiment ("Konus" is derived from the Russian for "cone" and refers to the spacecraft's configuration), four Soviet Venera space probes were equipped with extremely sensitive detectors. The experimenters, E. P. Mazets, S. V. Golenetskii and their colleagues at the A. F. Joffe Institute of Physics and Technology in Leningrad, used these detectors to compile a catalogue of the times, positions, spectra and variations in brightness of hundreds of separate bursts.

Although these more recent observations have determined the position of some bursts, it is difficult to acquire more specific information about any single burst. Because they are so brief and because they occur in unpredictable locations, it is impossible to find them using a sensitive detector with a narrow field of view. Instead, to have a practical hope of detecting any bursts at all, an observer must use a relatively insensitive, wide-angle detector.

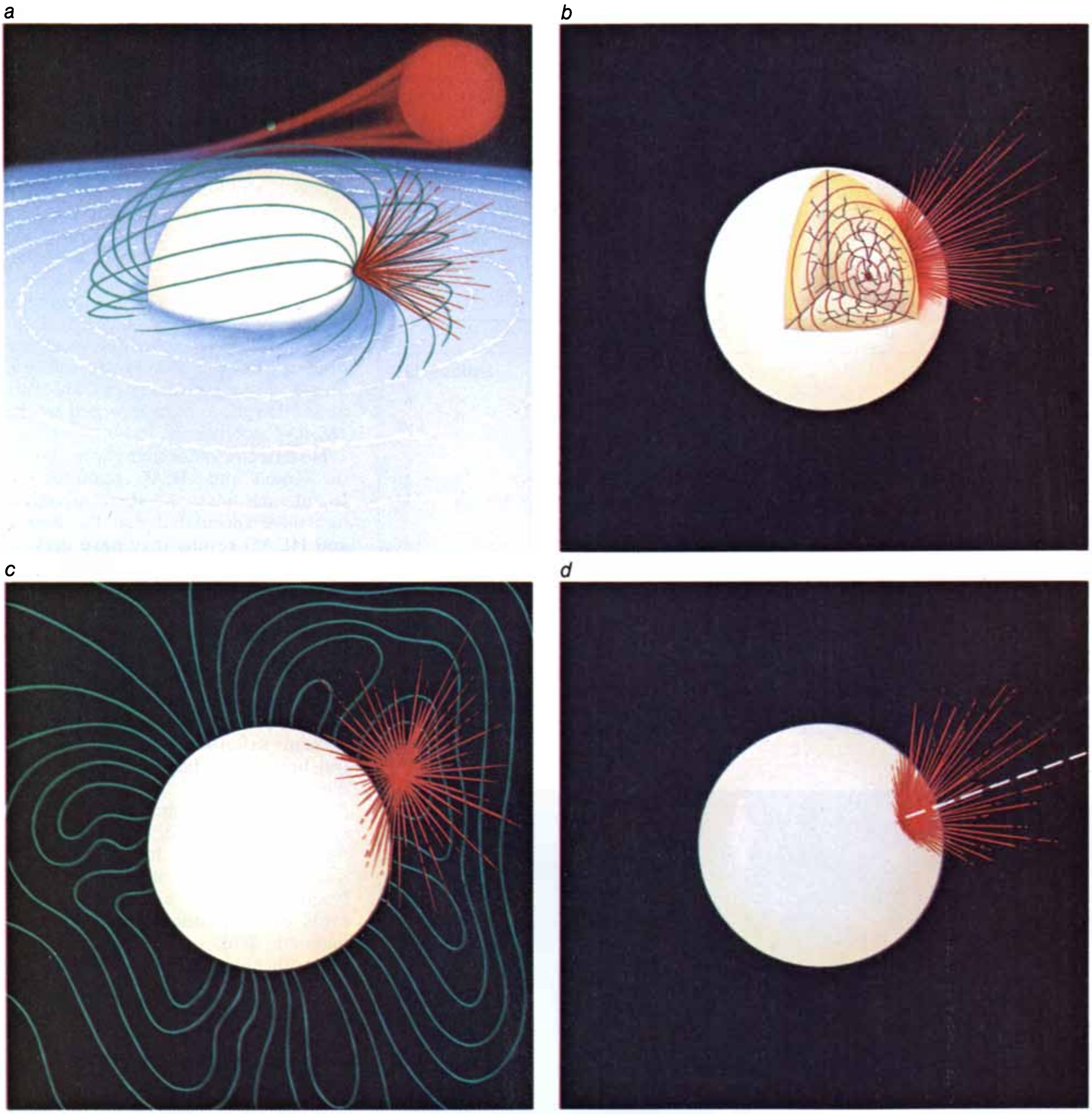
Further difficulties arise because the bursts consist almost entirely of gamma radiation, which is absorbed by the earth's atmosphere. Detectors must therefore be flown in space, an expensive and complicated process. In addition gamma-ray detectors are inherently less sensitive than detectors of radiation at lower energies.

In the face of these experimental difficulties, what kinds of data are there that could make it possible to distinguish among models? Fundamentally all gamma-ray observations consist solely of "light curves," records of the variation in brightness of a burst. Since each burst consists of radiation at several levels of energy, different light curves (one for each energy level) must be recorded for each burst. So far light curves have been collected for several hundred bursts, and curves for a single burst have often been recorded by more than one satellite. This small collection of light curves constitutes almost the entire body of our knowledge about gamma-ray bursters.

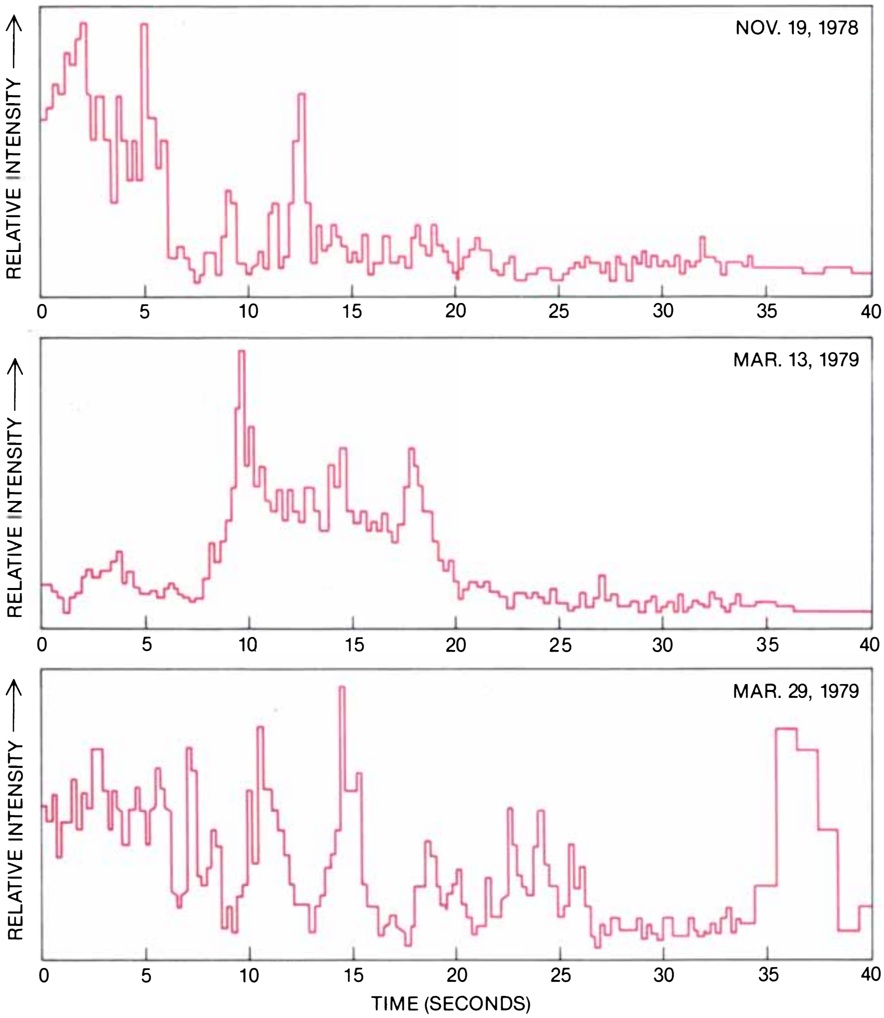
Meager as this may seem, there are ways of extracting useful information from light curves. For example, by comparing the intensities and times of curves recorded by detectors on different satellites, investigators can determine a burst's position in the sky with some accuracy. It is also possible, by comparing the light curves of different energy levels within the same burst, to determine its spectrum: the relative intensities of the different energies of radiation that make it up.

The shape of an object's spectrum can often reveal information about the nature of that object, such as its temperature and size. Many investigators have examined the observed spectra of the bursts in attempts to identify the emission process that causes them. Although several different radiation processes are known that would generate spectra like those observed for gamma-ray bursts, no one of them can satisfactorily explain the spectral shape of all the bursts.

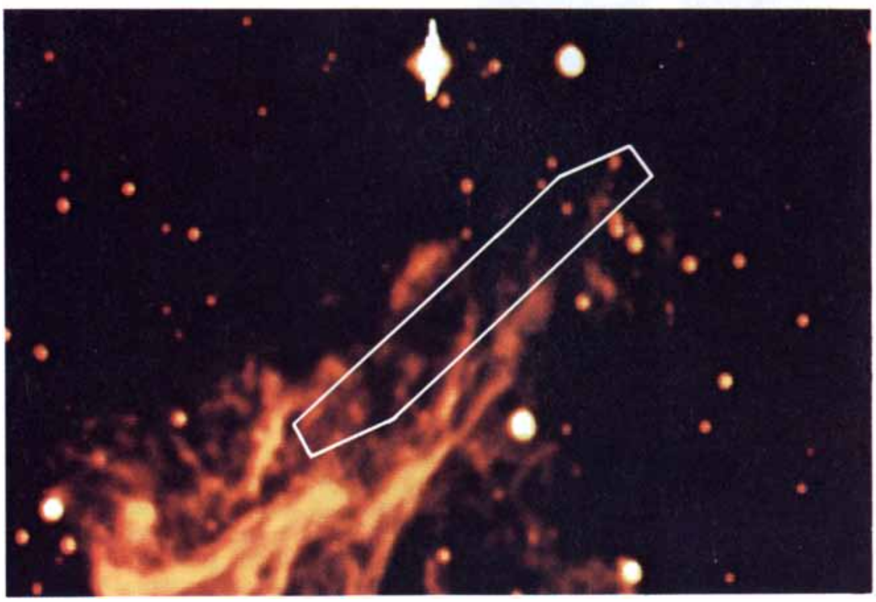
Indeed, it would be surprising if any



MODELS OF GAMMA-RAY BURSTERS have proliferated owing to the paucity of observational data that could provide a definitive test. The likelier models, four of which are shown, are based on neutron stars. In *a* some matter from a relatively faint companion star (*red*) gathers in a disk centered on the neutron star. It is then pulled along the neutron star's magnetic field lines (*indicated in green*) toward the magnetic pole. As the matter accumulates at the pole it is compressed and heated. Eventually it explodes, emitting gamma rays. In *b* the star experiences a "corequake." The quake spreads to the surface, where shock waves produce gamma rays. Another model (*c*) attributes the bursts to irregularities in the neutron star's magnetic field: very irregular field lines can become unstable and form a closed loop above the surface of a star. The loop shrinks. Ultimately it will disappear, and the large amount of energy that had constituted the irregular portion of the magnetic field will be released as electromagnetic radiation, some of which will be in the form of gamma rays. Bursts could also be caused by the impact of a comet or an asteroid (*d*) on the surface of a neutron star (*path in white*). Because of the star's intense gravitational field, such a collision would be energetic enough to produce gamma rays.



LIGHT CURVES show how the brightness of a burst varies over time. These three curves, which were produced by the Soviet "Konus" detector, describe three different bursts. They have no consistent structure. This is a frustrating circumstance, because light curves are among the very few unambiguous observational data available for gamma-ray bursts.



POSITION of the unusual burst occurring on March 5, 1979 (*outlined*), was determined with great accuracy. Burst was seen by 11 detectors widely spaced within the solar system; its position was found by comparison of times at which it was observed by various detectors.

of these theoretical mechanisms were able to explain the burst spectra. Each mechanism is valid only under idealized conditions—a single temperature, magnetic field orientation and intensity, and instant of time—and the true situation probably bears a complicated relation to those depicted in the idealized models.

Even if spectral information is not particularly useful in determining the properties of the region that emits the radiation, it may still hold information about the volume of space surrounding the emitting region. For example, if that region were to absorb or emit radiation at one of the frequencies contained in the burst, then a line would appear in the burster's observed spectrum. In fact, the Soviet Konus experiment found just such lines, usually within the range of 40 to 70 KeV (thousand electron volts), in roughly 15 percent of the bursts it detected. Independent evidence for the existence of these lines was found by Geoffrey J. Hueter of the University of California at San Diego, in data recorded by the *HEAO-1* satellite.

No detector other than those aboard the Konus and HEAO satellites has found such lines, and some investigators have speculated that the Konus and HEAO results may have derived from errors in calibration. Indeed, it is difficult to calibrate the efficiency of a detector in the 40-to-70-KeV range because of a complicated response due to iodine in the instrument. Detailed analysis of the Konus and *HEAO-1* experiments indicates, however, that they are probably correctly calibrated. It seems safe to assume that the spectral lines are in fact contained within the bursts.

The most popular hypothesis concerning the origin of the lines is that they are due to "cyclotron emission," radiation produced at sharply defined frequencies when electrons perform rapid circular motions in an intense magnetic field.

Another explanation is that the burst radiation is due to two sources very close together, each producing a different spectrum. In this picture the angular separation of the two sources, as seen from the earth, would be so small that the radiation coming from them would appear to come from a single source. It is possible that two different spectra could, when added together, produce a spectrum having such lines.

Whereas spectral information can be used to determine the mechanism behind the bursts, information about the intensities of the observed bursts can be used to find whether the sources are distributed isotropically (evenly) throughout space. One method is to plot a graph of the number of bursts that are more intense than any given level of brightness [*see illustration on page 57*]. When the points are plotted on a logarithmic scale, they should fall on a straight line with a particular slope if the bursts are distributed evenly. If many of the gamma-ray bursters lie at one particular distance from the earth, the graph will show an irregularity at the intensity corresponding to bursters observed at that particular distance.

At low intensities the graph does indeed curve away from the straight line, although there are too few data points to be sure of the statistical significance of this deviation. The easiest explanation for the observed shape of the curve is that nearby bursters are evenly distributed but distant bursters are clumped in space in some way, leaving large regions of space with few, if any, bursters. Such a situation could occur naturally if bursters are distributed in a disk-shaped region of the galaxy. Unfortunately many other types of distribution would also satisfy the data.

Furthermore, the curve is apparently inconsistent with our knowledge of the positions of observed bursts. The curve implies that distant bursters are not evenly distributed throughout space. Hence they should appear concentrated in some direction of the sky (such as that of the galactic plane or some nearby cluster of galaxies). As Mark Jennings of the University of California at Riverside observed in 1981, however, the faint bursts seem to be evenly distributed. To date five acceptable yet mutually exclusive explanations have been advanced as possible solutions to this problem.

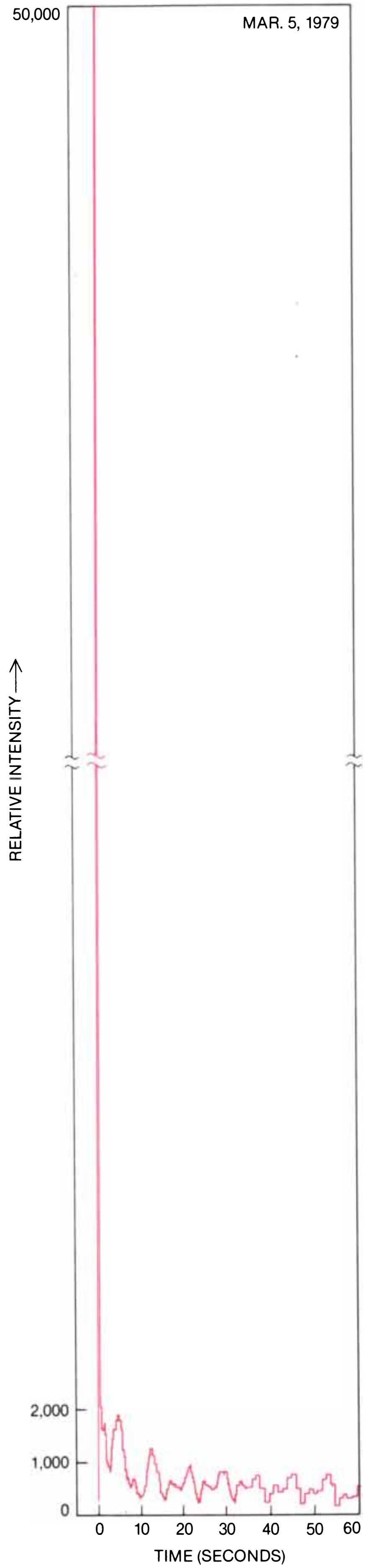
So far the picture I have painted is a grim one. Much of the basic observational information is flawed by inadequacies inherent in the detection apparatus. Even if the data are accepted, they are invariably open to many different explanations. This raises the question of whether any nontrivial facts are known concerning the nature of gamma-ray bursters.

Fortunately the situation may not be as bad as I have made it appear. In recent years investigators have reached a consensus about certain properties of gamma-ray bursters. It is centered around three basic points: that a gamma-ray-burst system contains a neutron star, that this neutron star has an intense magnetic field and that most of the observed bursters are situated within our own Milky Way galaxy. Although none of these points has been proved conclusively—and reasonable disagreement could be raised against each point—together they provide the simplest explanation for the available evidence.

The keystone of the recent consensus is the conviction that a neutron star is somehow involved in the burst of gamma rays. A neutron star is a very small and dense star formed during the later stages of stellar evolution. (The radius of a neutron star is roughly six miles, whereas its central density can be greater than 1,000 million tons per cubic inch.) The extremely large surface gravity and magnetic field of a neutron star certainly hold enough energy to generate a burst of gamma rays, and many mechanisms have been proposed that would convert energy from these sources into gamma radiation. Another reason theorists are attracted to models that include neutron stars is that neutron stars are known to exist and are relatively common entities in our galaxy.

These theoretical reasons show it is plausible that a gamma-ray burster might contain a neutron star; certain observational facts make it probable that it does. One such fact is the very short time within which bursts change their intensity. Some bursts have been as short as .01 second, whereas a burst that occurred on March 5, 1979, rose in intensity in .0002 second. Since a source cannot significantly change brightness in a time shorter than the time it takes light to travel across the source region, the size of the March 5 burster must be smaller than .0002 light-second, or about 40 miles. There are few astronomical objects that meet the size limitations or have enough available energy to power a burst. A neutron star satisfies both of these requirements.

Another observational argument for the involvement of neutron stars in gamma-ray bursts is based on a peculiarity found in approximately 7 percent of the burst spectra. These spectra have a line at roughly 420 KeV. There are few processes that could emit a prominent line at that energy; the most plausible hypothesis is that the lines are formed when electrons meet their antiparticles, positrons. Such a collision would result in the total annihilation of both particles and the conversion of their mass energy into two



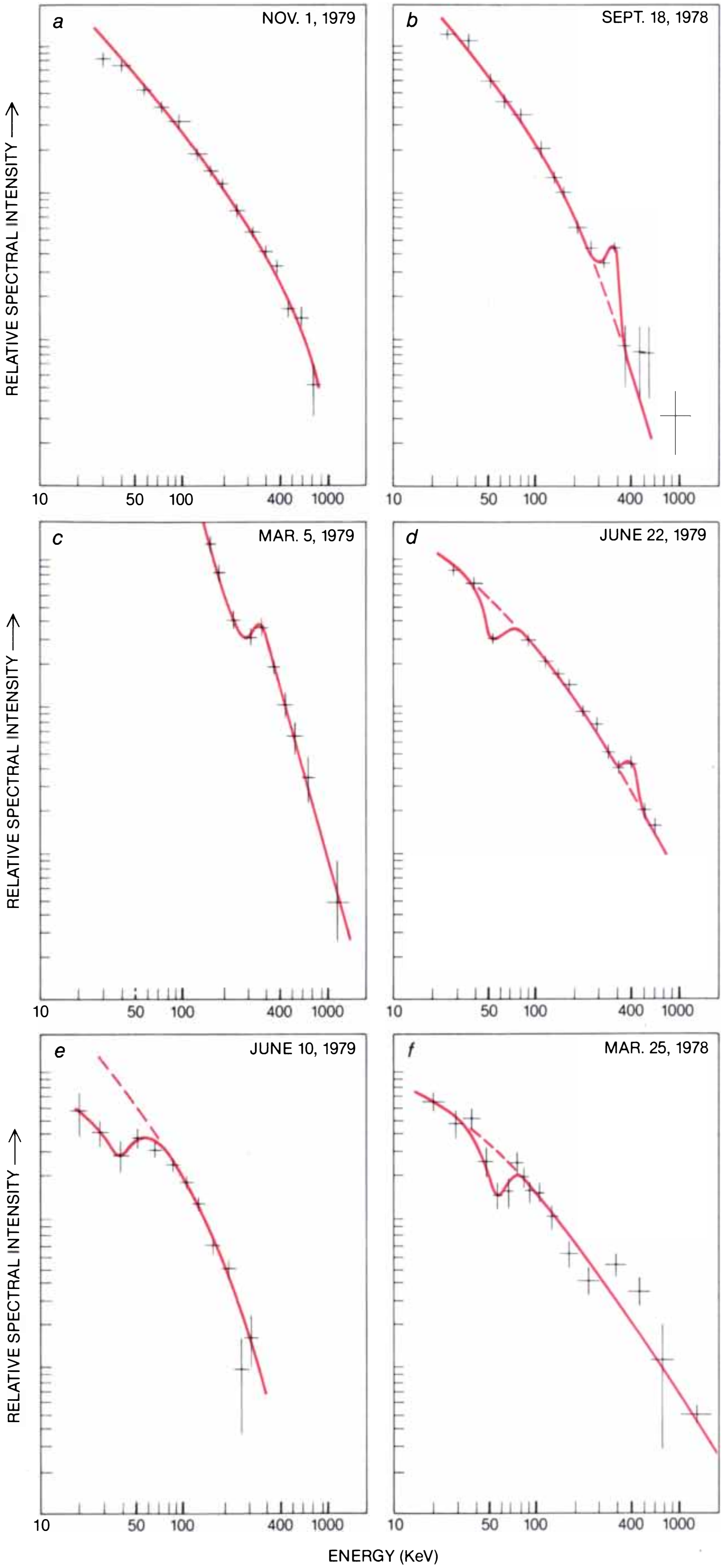
UNUSUAL FEATURES of the March 5 burst are apparent in its light curve. Its intensity rose in less than .0002 second to a peak more than 10 times as bright as any other known burst. As the burst decayed its intensity oscillated with a period of roughly eight seconds. This is the only established case of periodicity in a gamma-ray burst.

gamma rays. Each gamma ray would normally have an energy of 511 KeV; if this mutual annihilation took place near the surface of a neutron star, however, then before the gamma rays could reach the earth they would first have to emerge from the neutron star's gravitational well. In doing so they would lose energy. (This loss of energy is the so-called gravitational red shift predicted by the general theory of relativity.) In fact, the gamma rays would lose just enough energy to make up the difference between the 511 KeV at which they were emitted from the electron-positron collision and the 420 KeV at which some gamma-ray bursters show emission lines.

Further evidence for the idea that neutron stars are involved in gamma-ray bursters is provided by two unique characteristics of the March 5, 1979, burst. One of these characteristics is that the burst's position in the sky was very near that of a supernova remnant. If this association is taken at face value, the source of the burst is probably related to the remains of a supernova. Since neutron stars are often created by supernova explosions, it is reasonable to suppose a neutron star is at least partly responsible for the March 5 burst.

A second unique characteristic of this burst was that its brightness oscillated with a period of approximately eight seconds. Many periodic astronomical emissions can be explained by the rotation of a star: as the star rotates, a beam emitted by one area of the star's surface periodically strikes the earth. An eight-second rotation period is too fast for most types of star, but it is typical for neutron stars.

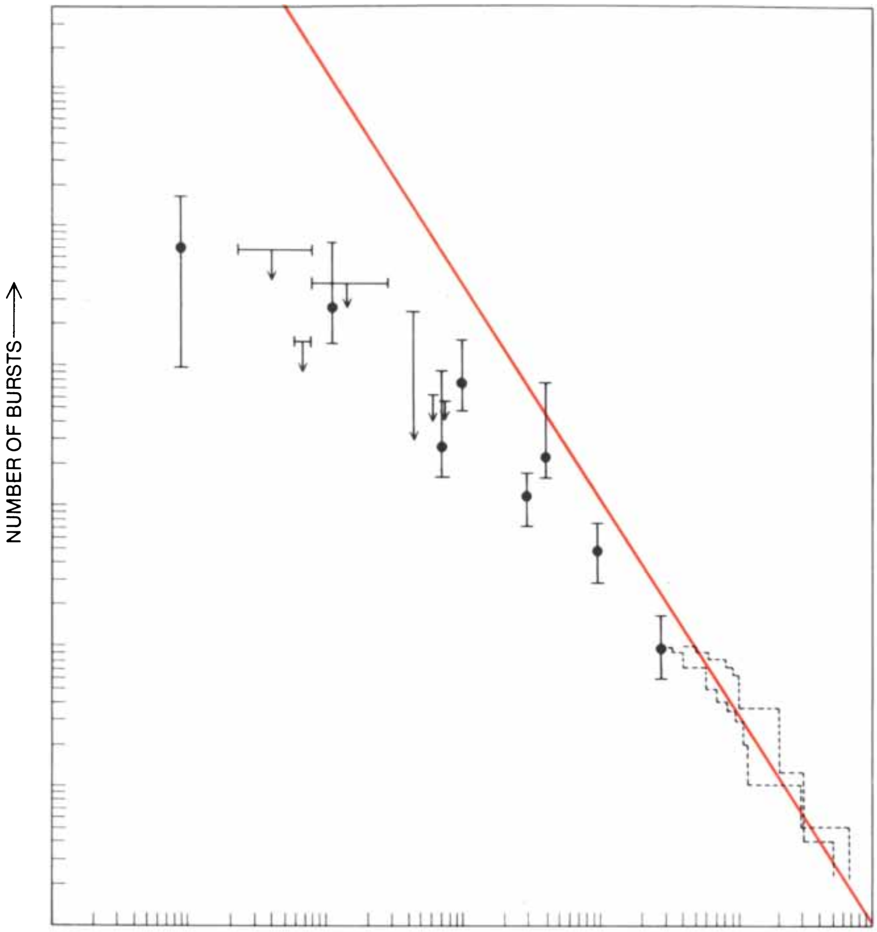
The second feature of the recent consensus on models for gamma-ray bursters is that an intense magnetic field is part of the phenomenon. Again, none of the evidence in support of this hypothesis is convincing in its own right; each item can be interpreted in a manner that does not require a magnetic field. In sum, however, they provide a reasonable basis for the inclusion of a strong magnetic field in the theory of gamma-ray bursters.



SPECTRA of gamma-ray bursts provide clues to the nature of the burst phenomenon. Some bursts show absorption or emission phenomena between 40 and 70 KeV (*d*, *e*, *f*). These may be due to electrons that perform rapid circular motion in an intense magnetic field near the source of a burst. Another prominent feature of some spectra (*b*, *c*, *d*), an emission line near 420 KeV, could be due to collisions of positrons and electrons.

One of the strongest of these arguments relies on the observation of emission or absorption lines in the region of 40 to 70 KeV. If these lines are indeed due to the circular motion of electrons, the magnetic field forcing the electrons to follow circular paths must be on the order of 10^{12} gauss. (In comparison, the earth's magnetic field is roughly half a gauss.) If the field were much smaller, the electrons would emit cyclotron radiation at a lower energy.

A second observational argument is based on the eight-second periodicity seen in the burst of March 5, 1979. This modulation was probably caused by the rotation of a neutron star; some emitting region of the star rotated periodically into and out of the earth's field of view. Such a region could not lie on one of the poles of the star, where the axis of rotation meets the star's surface, or it would not rotate and hence would not pass through the earth's line of sight. There must therefore be some mechanism, not symmetric about the axis of rotation, that determines the location of the emitting region on the star's surface. A likely candidate for such a "symmetry breaking" mechanism is a magnetic field whose axis is misaligned with the rotation axis of the neutron star.



SPATIAL DISTRIBUTION of the bursters can be gauged by plotting the number of bursts that are brighter than a given intensity. If the bursters are evenly distributed, the curve (when plotted on a logarithmic scale) should follow a straight line with a slope of $-3/2$ (*color*). Curve's deviation from this line indicates that observed bursters may not be uniformly distributed. Broken lines represent catalogues of data compiled by individual satellites.

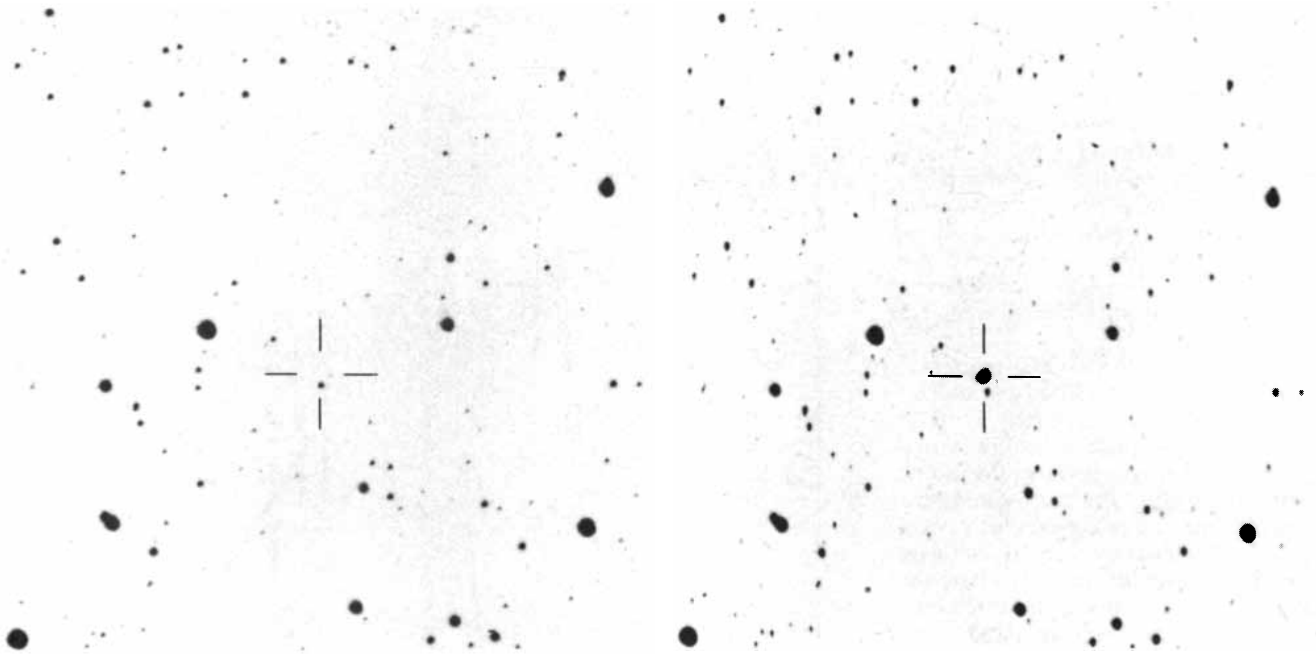
The hypothesis that a magnetic field is present in a gamma-ray burster is supported by theoretical as well as observational arguments. One of the strongest of these is that any region that emits intense gamma radiation must be confined by some force; otherwise radiation pressure generated by the gamma rays would cause it to expand very rapidly. If the region expanded too far, its density would become so low that it would no longer emit gamma rays. Even the tremendous surface gravity of a neutron star is not adequate to confine an intense source of gamma rays, but a magnetic field of about 10^{12} gauss would probably be sufficient.

The third feature of the recent consensus is that most of the gamma-ray bursters are in our own galaxy. One of the strongest arguments for this hypothesis is based on the great amount of energy the bursters would have to produce if they were extragalactic. A burster in the Milky Way would have to generate roughly 10^{38} ergs of energy to achieve the brightness observed on the earth; an extragalactic burster would have to generate at least 10^{46} ergs to achieve the same level of brightness. Whereas it is relatively easy to devise models of processes that would generate 10^{38} ergs, there are many serious difficulties in producing a model that would generate as much as 10^{46} ergs.

An observational argument, based on the graph of the number of bursts brighter than certain intensities, also supports the hypothesis that bursters are within the Milky Way galaxy. If most of the observed bursters were outside the galaxy, they would be most likely to exist within other large accumulations of mass (for example other galaxies or clusters of galaxies). The observed distribution of bursters in space would therefore be irregular. Large concentrations of them would fall at various distances from the earth. In that case the graph would show a series of irregularities, corresponding to the distances to various concentrations of bursters. The curve shows no such irregularity.

Although the current consensus is a good starting point, its theoretical usefulness is limited by the difficulty of obtaining good data from bursts of high-energy photons such as gamma rays. Astrophysicists have realized for a long time that many observational difficulties could be solved if there were some way to observe lower-energy photons coming from the bursters. To this end much effort has been spent in determining accurate positions for as many bursts as possible. These specific positions have then been examined with optical telescopes, to see if any optical phenomena are associated with the bursts.

With two exceptions the results of this search have been consistently negative. The first exception was the unusual burst of March 5, 1979, which has been tentatively identified with the remains of a supernova. The second exception consists of three optical flashes, which I found in 1981 while I was a graduate student at the Massachusetts Institute of Technology. In my search for optical flashes from gamma-ray bursters I examined ap-



OPTICAL FLASH appears in one of two photographs made on November 17, 1928; the photograph on the left was made 45 minutes earlier than the one on the right. This flash occurred in the same region of the sky from which a gamma-ray burst emanated on November 19, 1979. Star images in the right-hand photograph are elongated (trailed) because the telescope did not track the sky perfectly. The flash itself does not show such trailing, indicating its duration was significantly shorter than the total exposure time.

proximately 30,000 photographs, part of the archival collection of 500,000 photographs owned by the Harvard College Observatory. I searched each of the archival photographs in Harvard's collection that recorded an area of the sky where a gamma-ray burst had later been observed. In three photographs (of three separate regions), taken in 1901, 1928 and 1944 respectively, I found starlike images that did not appear in other photographs taken of the same regions. These images, which were caused by optical flashes, are certainly related to the gamma-ray bursts that were observed later in the same locations.

Although it would be useful to have more data, much can be learned from these three flashes alone. For example, in all three cases the energy contained in the recent gamma-ray burst is almost precisely 1,000 times the energy contained in the optical burst. Taken together with various theoretical arguments, this ratio can be used to show that the neutron star causing the bursts must have a companion of some kind, perhaps a faint star or an accretion disk of colder matter.

Since the positions of the bursters can be determined more accurately from optical data than from gamma-ray data, the three optical positions have enabled us to make deep searches for quiescent bursters, using large optical telescopes. Although the results are not conclusive for these three positions, it is apparent that bursters must be extremely faint when they are not bursting.

Statistical analysis of the three optical flashes shows that gamma-ray bursters must emit such flashes at an average rate of once per year: I examined an average of three months' worth of archival photographs for each of 12 positions known to contain gamma-ray bursters—in effect, I examined a cumulative exposure of three years. In this three-year collection of photographs I found three flashes, which indicates that the average time between optical flashes of a gamma-ray burster is approximately one year. This result may preclude those models that do not allow the possibility of yearly optical flashes.

Perhaps the most important implication of these three flashes is that they have forcefully reminded astrophysicists that it is both practical and desirable to observe gamma-ray bursters at lower energies.

The next few years will see the initiation of several experiments that have great potential for acquiring new information on gamma-ray bursters. The most exciting of these experiments is the burst detector on the Gamma Ray Observatory satellite, scheduled for launching in 1988. This instrument will not be fundamentally different from earlier detectors, except that it will be many times more sensitive. The Gamma Ray Observatory instrument will detect bursts 10 to 100 times fainter than the Konus experiment could have, and it will position these bursts on the sky to within one degree. In addition a special modification of the detector will record spectra over a wide range of energies for each observed burst.

A second major experiment, now in the final stages of construction, is a search for optical flashes. It will consist of two parts. The first, constructed by Roland Vanderspek and George R. Ricker, Jr., of M.I.T., will monitor most of the visible sky for any sudden brightening. Within a second of the start of any flash its position will be relayed to the second part of the experiment, an optical telescope that can point to any region of the sky within one second. (This part of the experiment is being built by Bonnard J. Teegarden, Ravi Kaipa, Tycho T. von Rosenvinge and Cline at the Goddard Space Flight Center.) The investigators expect to detect roughly two dozen bursts per year.

The results we expect from these and other experiments should finally reveal the underlying cause of gamma-ray bursts. In spite of the current consensus, the burst phenomenon remains one of the most mysterious of all classes of astrophysical events. I look forward to the day when data will be available that will enable us to understand it.

SCIENCE AND THE CITIZEN

SALT Shakers

Can the Soviet Union be trusted not to violate an arms-control treaty? Now that delicate efforts are under way to get the two superpowers back to the negotiating table, the question is often asked—sometimes rhetorically.

The Reagan Administration itself has put the question in the foreground. In January of 1984 the Administration issued a report to Congress listing seven Soviet "violations and probable violations" of arms-control agreements; an update of the report is due to be published this month. There is also a classified report prepared by the president's General Advisory Committee on Arms Control and Disarmament; the secret report was the subject of much publicity during the recent presidential elections, and its summary, which was declassified in October, lists numerous "material breaches" as well as "suspicious events" and "breaches of the duty of good faith" dating back to events preceding the Cuban missile crisis. The Soviet government has responded to these reports by producing lists of alleged U.S. treaty violations.

To what extent is it true that the U.S.S.R. (or the U.S.) has a poor record of compliance, and what are the implications of such "cheating" for the future of arms-control efforts? The first of these questions is not as straightforward as it may seem. Because many of the terms and conditions of the Strategic Arms Limitation Talks (SALT) treaties are ambiguous and open to interpretation, it is rare to find an action by either the U.S.S.R. or the U.S. that clearly violates some provision of a treaty.

One example of this kind of ambiguity involves a large phased-array radar (a very precise, electronically steered radar) now being constructed by the U.S.S.R. near the city of Krasnoyarsk in central Siberia. The U.S. says this radar may be in violation of the anti-ballistic-missile (ABM) treaty, which prohibits either side from deploying early-warning radars except outward-looking ones near national borders.

The U.S.S.R. counters that the Krasnoyarsk radar will track satellites and is therefore allowed under a section of the treaty that reads: "The Parties agree not to deploy phased-array radars... except for the purpose of tracking objects in outer space...." It is thus unclear whether the Krasnoyarsk radar is in violation: although it could be used for certain space-tracking duties, it is not very well situated or oriented for such a purpose. U.S. arms-control experts believe the Krasnoyarsk radar may represent an infringement of the spirit, if not the letter, of the ABM treaty. As one analyst puts it, "the Soviets push at the margins of an agreement." That is, he explains, they take actions so close to the treaty limits that it is difficult to tell whether or not a breach has occurred. Analysts speculate that the U.S.S.R. may be testing the limits of U.S. intelligence-gathering capability.

An example of this kind of testing may be the Soviet practice of encrypting some of the telemetry transmitted by prototype missiles during tests. Because a provision of SALT II forbids deliberate concealment of data that could be used to verify compliance with the treaty, the U.S. has contended that Soviet encryption violates the agreement. (Although the U.S. has not ratified SALT II, both sides have provisionally agreed to abide by its conditions.) The U.S.S.R. responds to this charge by asking what additional data are needed in order to verify that testing of the missile complies with the treaty. There dialogue ends, analysts say, because any reply to this request would compromise U.S. intelligence sources and methods.

Another U.S. charge is based on a provision of SALT II that limits each side to one new type of intercontinental ballistic missile (ICBM). The U.S. maintains the U.S.S.R. has tested two new types of missile, the SS-X-24 and the SS-X-25; the U.S.S.R. has replied that the SS-X-25 is actually a modification of the older SS-13 missile. Under SALT II an existing ICBM may be modified but its throw weight (the weight of its warheads, the "bus" that distributes them and any penetration aids they carry) may not change by more than 5 percent. The U.S. asserts the SS-X-25 throw weight is larger than this rule allows, but the U.S.S.R. answers that U.S. intelligence data on the old SS-13 are faulty. The Administration responds that even if the SS-X-25 is not a new missile, it violates another term of SALT II that prohibits testing an ICBM with a warhead smaller than half the missile's throw weight.

For its part the U.S.S.R. offers a list of U.S. violations. At the top is a new American phased-array radar called PAVE PAWS [see "Phased-Array Radars," by Eli Brookner, page 94]. While all PAVE PAWS installations are on the periphery of the U.S., the radar has an angle of view wide enough to enable two installations under construction, one in Georgia and the other in Texas, to provide coverage of as much as two-thirds of the continental U.S. The U.S.S.R. has charged such coverage may violate the ABM treaty.

The question of PAVE PAWS compliance, like many of the compliance questions raised by the U.S.S.R., concerns steps initiated, but not yet completed, by the U.S. Two other examples are the so-called Midgetman missile, which would be the second new type of ICBM tested by the U.S. after SALT II, and the "Star Wars," or Strategic Defense Initiative, space-based missile-defense system; development of this system could be a violation

An ABM-treaty violation? The Krasnoyarsk radar is not on the Soviet border of the ABM treaty. The Soviet Union has also protested some actions already taken by the U.S. One such case involved shelters erected over Minuteman ICBM silos during refurbishing. The U.S. said the shelters were to protect workmen from the weather; the U.S.S.R. said they impeded verification and were in violation of SALT I.

In order to resolve such issues SALT I provided for a bilateral Standing Consultative Commission (SCC), a confidential forum in which delegations from each side raise concerns they have regarding the other side's compliance. A typical Soviet reaction is to deny that the activity in question violates an agreement and then to desist quietly. Herbert Scoville, Jr., a former deputy director of the Central Intelligence Agency and of the Arms Control And Disarmament Agency, says it is important that American concerns be raised in a manner that makes such diplomatic correction possible: "Once you accuse somebody, they all get their backs up.... Nobody likes to admit they have done something questionable, and therefore they don't dare stop it."

"It is not important to have courtroom evidence," says Robert W. Buchheim, who was the U.S. SCC commissioner (head of the U.S. delegation) from 1977 to 1981. "We needn't look at the fine print to be able to say we find [a possible infraction] offensive." At that point, he observes, the two sides can work toward a solution that preserves the integrity of the agreement.

Many of those close to the arms-control process stress the importance of mechanisms such as the SCC in maintaining strategic stability. Albert Carnesale of Harvard University's John F. Kennedy School of Government points out that even such potentially serious violations as the Krasnoyarsk radar would not, taken alone, provide the Soviet Union with the ability to "break out" of a treaty with a significant military advantage. The function of the SCC is to keep possible violations and misunderstandings at such a low level that they do not threaten the military balance. Gerard C. Smith, chief negotiator of the SALT I and ABM treaties, adds: "Our chances of getting corrections of these [possible infractions] is greater in an ongoing arms-control context than if we are just snarling at each other."

No Taps for TTAPS

Ever since the nuclear-winter hypothesis was first put forward more than a year ago by a group of American atmospheric physicists and planetary astronomers, it has been severely criticized on both scientific and political grounds. Emphasizing the many uncertainties inherent in any attempt to calculate the long-term climatic effects of nuclear war, some critics have suggested that it was improper for scientists even to discuss the issue in public before the technical basis for such a dire forecast could be better established. Others have gone further, imputing a political motive to the authors of the original nuclear-winter scenario, known as the TTAPS study (after an acronym formed from the authors' names: Richard P. Turco, Owen B. Toon, Thomas P. Ackerman, James B. Pollack and Carl Sagan).

In response to such criticism the TTAPS group has argued essentially that they are well aware of the scientific uncertainties of their subject, that they took great pains to have their initial calculations reviewed by recognized experts in the relevant fields and that in any case their paper clearly expressed the hope that "the scientific issues raised here will be vigorously and critically examined." Pointing out that nuclear winter is after all "not a subject amenable to experimental verification—at least not more than once," the group has held that "open and informed debate on this issue is the only responsible approach, given the gravity of the potential climatic catastrophe we believe we have uncovered."

The latest indication that the scientific community at large has taken the nuclear-winter hypothesis seriously was the publication in December of a 193-page report by an interdisciplinary panel assembled by the National Research Council to review the available evidence. The NRC report, which was undertaken at the request of the Department of Defense, concluded that although it is impossible at this time to give "accurate detailed accounts" of the climatic effects of nuclear war, owing to the large uncertainties in current knowledge, there is nonetheless a "clear possibility" that the nuclear-winter model is valid. Echoing the call of the original TTAPS paper, the NRC panel recommended that efforts to reduce the uncertainties still surrounding the possible atmospheric effects of a major nuclear war "should be given a high priority."

Earth Moves

According to the theory of plate tectonics, the lithosphere—a layer some 100 kilometers thick consisting of the earth's crust and part of the mantle—is divided into rigid plates that constantly move. Such a plate travels outward from a midocean rift, where upwelling magma adds new lithosphere to the trailing edge; simultaneously, hundreds of kilometers away, the leading edge of the plate plunges into the earth's interior at a seismically active region known as a subduction zone.

Two important questions emerge from this picture. Why do new rifts, which split a single plate in two and open a new ocean basin, almost always develop on land rather than in a preexisting ocean? What explains the fact that earthquake activity varies from one subduction zone to another? Writing in the *Journal of Geophysical Research,* geophysicists from Princeton and Stanford universities and the International Institute of Seismology and Earthquake Engineering in Japan have provided convincing answers.

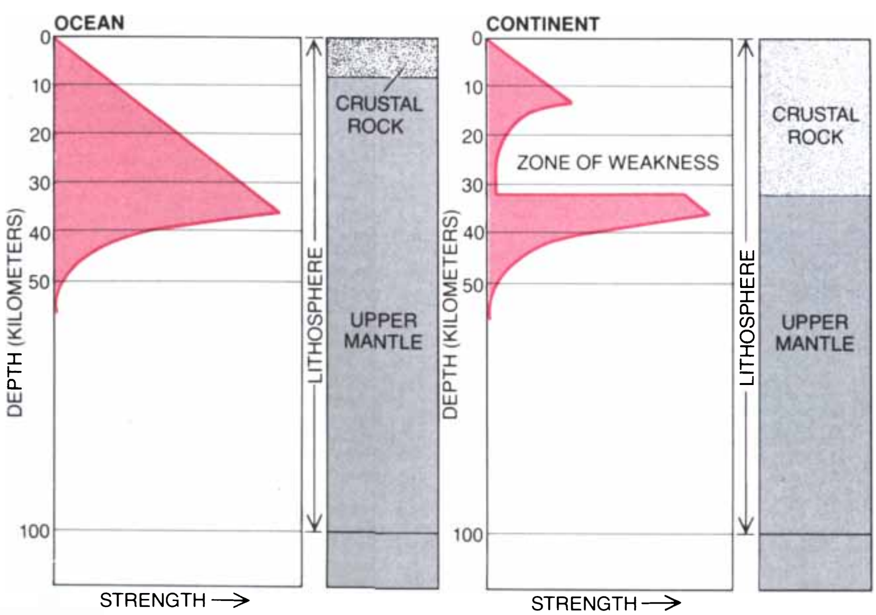
The Princeton workers, Gregory E. Vink, W. Jason Morgan and Wu-Ling Zhao, have found that rifts open in continents rather than in ocean basins because the tensile strength of continental lithosphere is relatively low. Like a jelly sandwich, continental lithosphere has two strong layers and an intervening zone of weakness. This odd structure develops because in outer layers of crust, where the rock is brittle, strength increases with pressure and therefore with depth. At a depth of about 13 kilometers the picture changes abruptly: high temperature softens the rock and its strength decreases sharply.

Below this weak zone, at the base of the continental crust perhaps 30 kilometers down, the composition of the rock shifts; it becomes characteristic of the upper mantle. Since mantle rock remains brittle and strong even at the high temperatures found in this region, lithospheric strength rises once again.

In oceanic lithosphere the transition to mantle rock occurs at much shallower depths, typically five to 10 kilometers. Oceanic lithosphere therefore lacks an intermediate weak layer, and it increases continuously in strength down to a depth of 30 to 45 kilometers.

At the end of its career lithosphere created at a rift dives under another plate. Coupling, or binding, between the subducting plate and the overriding one often sets up stresses that lead to earthquakes. One would expect coupling and hence earthquake activity to vary with the rate at which the two plates converge, but Eric T. Peterson of Stanford and Tetsuzo Seno of the International Institute of Seismology and Earthquake Engineering could find no such relation.

Instead their data showed that in different regions of a single subducting plate seismicity is most intense where the subducting lithosphere is youngest. The authors point out that older lithosphere, which has had more time to cool since its volcanic genesis, is denser and sinks more readily into the mantle. Coupling is therefore reduced, they hypothesize, and with it the energy released as earthquakes.



Strength plotted against depth for typical oceanic and continental lithosphere

Factor Factory

Victims of classic hemophilia (hemophilia *A*) bleed because they lack a single crucial protein, factor VIII, implicated in the complex cascade of reactions that leads blood to coagulate. Prophylactic administration of a plasma concentrate enriched in factor VIII compensates for the defect. The treatment carries some risk. Because the plasma is ordinarily pooled from a large number of donors, it can be contaminated with blood-borne viruses, including the agents of hepatitis and AIDS.

Two biotechnology companies, Genentech and the Genetics Institute, have now cloned the gene for factor VIII, introduced it into mammalian cells in a laboratory culture and got the cells to synthesize biologically active factor VIII. In a few years a pure factor VIII manufactured in quantity by cultured cells may well be available to treat hemophilia.

Richard M. Lawn and Gordon A. Vehar led the Genentech team; John J. Toole and Rodney M. Hewick led the one at the Genetics Institute. Both groups report their work in *Nature*.

In recombinant-DNA technology the first step in making a protein is ordinarily to harvest the molecule's messenger RNA (mRNA): the nucleic acid that carries the genetic message from the DNA of the gene to the ribosomes, where amino acids are assembled to make proteins. The mRNA is then copied to make complementary DNA (cDNA), an artificial gene that directs the synthesis of the sought-after protein in bacterial or yeast cells.

This strategy was not available to the Genentech and Genetics Institute workers because factor VIII is extremely scarce and its mRNA is too hard to find. Instead the two groups began by purifying the factor VIII protein itself. They determined parts of its amino acid sequence, which they reverse-translated to derive corresponding short DNA sequences. Then they assembled off-the-shelf nucleotides (the subunits of DNA) to synthesize the short fragments of DNA. The fragments served as probes with which to find pieces of the actual, extremely long factor VIII gene.

To assemble the complete coding region of the gene, however, the two groups still needed to find a sufficient quantity of the intact factor VIII mRNA. Now they had the pieces of the actual gene to serve as probes with which they could identify the mRNA. The Genentech group found their mRNA in a human *T*-cell line; the Genetics Institute group used human liver cells. In each case the mRNA was copied into cDNA encoding factor VIII. The cDNA was combined with a viral "promoter" and inserted in a plasmid (a small circle of DNA).

Both groups introduced their recombinant plasmid into mammalian cells (cells from hamster or monkey kidneys), which seemed likely to have the proper enzymes for processing the gene and the protein. The cells did express the gene, making a protein that behaved like factor VIII in various laboratory tests and that reduced the clotting time of plasma taken from hemophilia patients. Testing in animals should follow within the year.

Printer's Mark

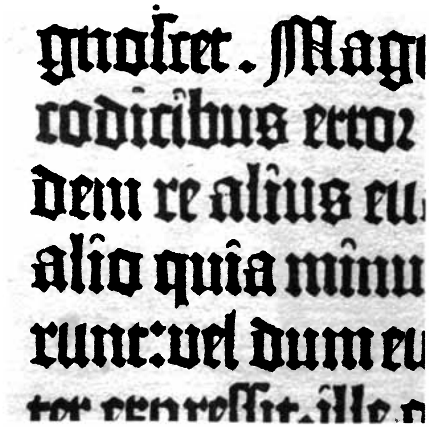
Sometime in the 1440's or 1450's Johannes Gensfleisch zur Laden, better known as Johannes Gutenberg, invented a method of printing from movable metal type. One component of this technological leap was an oil-based ink that, unlike earlier water-based preparations, would adhere to metal. Historians have long assumed that Gutenberg's ink was a mixture of linseed oil and lampblack, or soot, a formula known to have been common by the 16th century.

Richard N. Schwab and Thomas A. Cahill of the University of California at Davis have now shown the assumption to be wrong: Gutenberg employed a superior formula consisting largely of copper and lead. The workers made this discovery by analyzing most of the pages in one copy of Gutenberg's masterpiece, the 42-line Bible. The analysis involved exposing the pages to a proton beam in a cyclotron. Energetic protons excite molecules in the ink, causing them to fluoresce in the X-ray range; each element emits a characteristic spectrum, and the intensity of the radiation is a measure of the abundance of the element.

The use of metals instead of soot as a pigment was probably inspired by oil painting. It helps to explain why Gutenberg's print is still glossy and black after five centuries, whereas many later works have faded badly. Copper and lead oxides are stable, but the carbon in lampblack oxidizes to carbon monoxide and so tends to evanesce.

More important, the high metal content of his ink distinguishes Gutenberg from other printers of the period. It thus constitutes a chemical signature that may help to settle long-standing disputes about the attribution of early works. The Davis group has found that the ink in the rare 36-line Bible is heavy in copper and lead, as is that of the *Sibyllenbuch*, an early fragment. The *Catholicon*, a later work credited to Gutenberg by some experts, does not bear his metallic mark.

Nor do some of the pages of the 42-line Bible itself. Many of the 48 extant copies contain "variant" pages printed from remade plates when popular demand forced an expansion of the press run. Schwab and Cahill have found that the ink on some of these pages contains much less copper and lead. The finding may be evidence of the end of Gutenberg's career. Late in 1455—perhaps before the Bible was done—he lost control of his press in a suit filed by a financial backer. Forced into retirement, he may have taken the secret of his ink with him.



The print in Gutenberg's 42-line Bible

Water Gap

One line of thought about how the earth formed is that it accreted from smaller bodies, or planetesimals. The simplest hypothesis is that the planetesimals were homogeneous: the ones arriving first were identical in composition with the ones arriving last. Examining this hypothesis, Manfred A. Lange of the Alfred Wegener Institute for Polar Research in Germany and Thomas J. Ahrens of the California Institute of Technology have uncovered a difficulty. At a time when the prospective earth was about half its final size, its gravitation would have ensured that further planetesimals would arrive at high velocity—high enough so that water would be liberated from the accreting planet and so be free to react with other substances, releasing hydrogen gas. The gas could then escape, depleting the earth of its water forever. Things could not have happened that way.

Lange and Ahrens exposed the difficulty by examining two aspects of accretion. First, they determined by experiment the strength of impact necessary to free water from minerals, notably serpentine. It turns out that when the earth was half its final radius, the impact of further planetesimals would have freed all the water.

Second, Lange and Ahrens examined how the water would have reacted with iron as the earth accreted further. The problem is that all the water reacts. For example, water, iron and the mineral enstatite react to form forsterite, fayalite and hydrogen gas. Iron is abundant enough so that no free water is left. For its part the hydrogen escapes: the accreting earth's gravity is insufficient to hold it.

Lange and Ahrens resolve the dilem-

WHO DID WHAT?

(1) GTE (2) GRUMMAN (3) RCA (4) GE
(5) ITT (6) HUGHES (7) TRW (8) WESTINGHOUSE

Test your corporate I.Q. Look over the eight achievements pictured here and match each to the corporation responsible for it. Take your time; we'll wait.

Finished? O.K., let's review your answers. Who is building the first high-powered Direct Broadcast Satellite? Hughes? No, RCA.

Who supplied the video cameras, video equipment, and transmission services that sent back the stunning, live pictures from the Space Shuttle? Westinghouse? No, RCA.

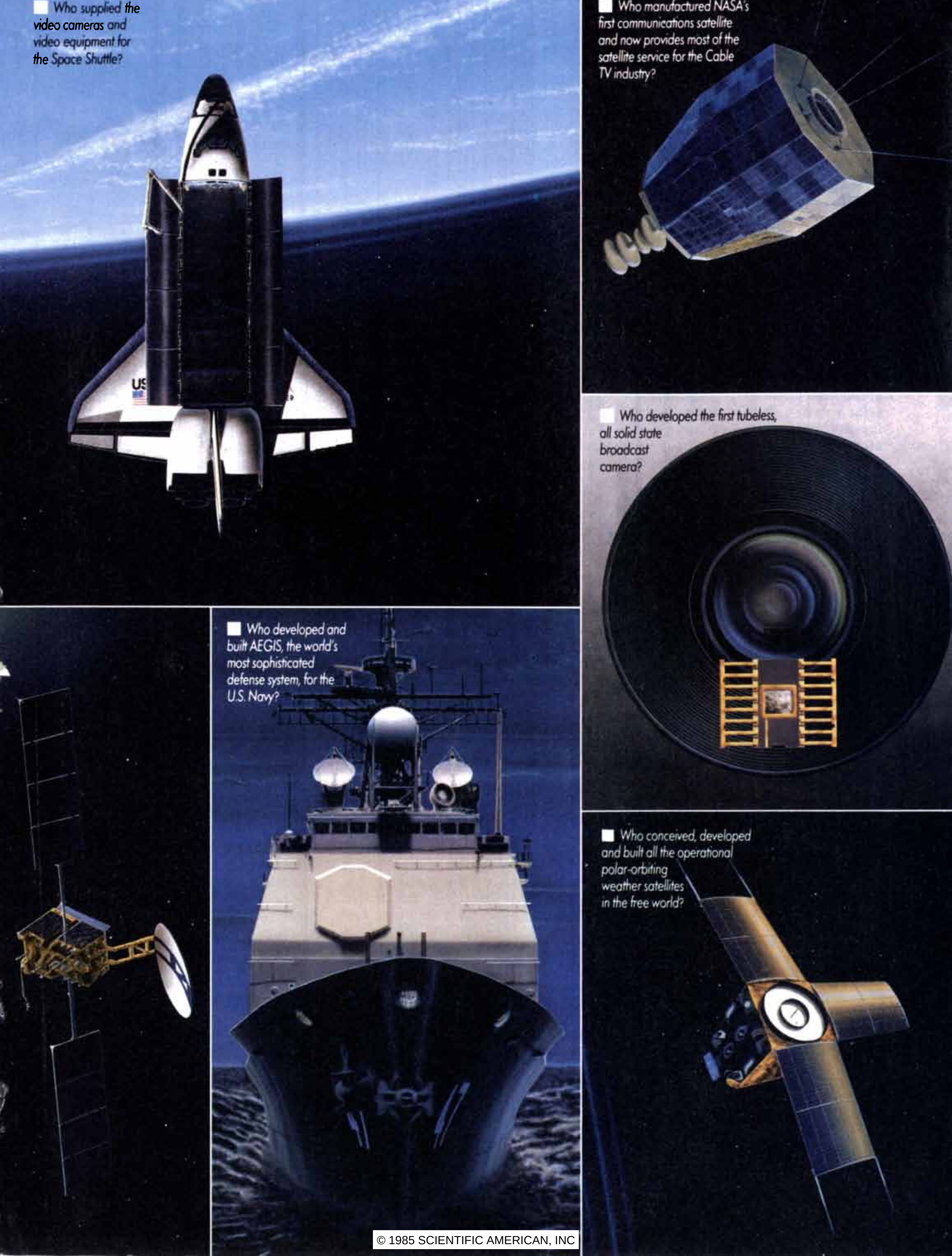
Who developed and built AEGIS, the Navy's new seaborne electronic defense system? Again, the answer, RCA. In fact, the correct answer to each of the questions is the same. RCA.

Surprised? If you're like most people, you aren't aware of what a major force RCA has become in many high-tech industries. Or that we invested over $200 million of our own money in R&D last year. A figure that compares with what many purely high-tech companies spend.

If you'd like to learn more about RCA as a corporation, an investment or as a place to work, write to: This Is RCA, P.O. Box 91404, Indianapolis, Ind. 46291. We'll send you a few more surprises.

In electronics, communications and entertainment, RCA is one of a kind.

RCA
ONE OF A KIND



Who supplied the video cameras and video equipment for the Space Shuttle?
Who manufactured NASA's first communications satellite and now provides most of the satellite service for the Cable TV industry?
Who developed the first tubeless, all solid state broadcast camera?
Who developed and built AEGIS, the world's most sophisticated defense system, for the U.S. Navy?
Who conceived, developed and built all the operational polar-orbiting weather satellites in the free world?

ma by positing an inhomogeneous accretion, in which the first planetesimals include more iron than the later arrivals. In the model that best produces the earth's current oceanic reservoir of 10^{25} grams of water, the first planetesimals are 36 percent iron by weight, the last have almost no iron and the average content of iron in all of them is 34 percent. Hence, over a period of about 100 million years, the earth accretes from planetesimals in an iron-rich environment. Then, over a period of about 60 million years, the environment is iron-poor.

Why did the planetesimals change? Kerry Harrigan and William Ward of the Jet Propulsion Laboratory propose that the solar nebula—the cloud of gas and dust that became the solar system—may have included density waves: spirals of high concentration of preplanetary matter. Conceivably the solar nebula's density waves nudged the developing planets outward. The accreting earth might thus have been moved from a region near the sun to a more distant, colder region, where the chief condensates would be silicates poor in iron.

Perennial Proton

More than 600 meters under the ground, in the cavity of a salt mine near Cleveland, Ohio, 2,048 photomultiplier tubes are suspended in a pool of 8,000 tons of water. They have waited there since July, 1982. If the photomultipliers fire in the proper sequence over a short time, their combined signal would be highlighted by a computer program. The computer would in turn alert a collaborating group of physicists from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory to a momentous and—on a cosmic scale—ominous event. The event is the decay of a proton [see "The Decay of the Proton," by Steven Weinberg; SCIENTIFIC AMERICAN, June, 1981]. If such a decay is ever detected, it would help to confirm a theory that links three of the four fundamental forces of nature into a grand unified theory, and it would imply that all matter as we know it will eventually disintegrate. No such decay has yet been confirmed, and the result suggests grand unified theory will require modification.

According to the simplest version of the theory, called minimal *SU*(5), the proton should decay into a positron and a neutral pi meson at least once, on the average, every 5×10^{31} years. The experiment in the salt mine is not expected to take that long; instead the 10^{33} protons making up the water in the chamber are monitored simultaneously, and about 20 proton decays are expected every year. The data so far suggest the average lifetime of the proton is much longer than the lifetime predicted by minimal *SU*(5): at least 1.2×10^{32} years.

The negative result does not entirely rule out other schemes for unifying the forces of nature. Several versions of grand unified theory based on mathematical groups more complicated than the group *SU*(5) remain consistent with the experimental data, primarily because such versions are unable to make any quantitative prediction about the proton's lifetime. Moreover, although the Irvine, Michigan and Brookhaven (IMB) group has examined some 30 possible channels of proton decay, it remains possible that other channels have still gone undetected or that data already collected will eventually be interpreted as a proton decay. The IMB group received approval recently for a more elaborate version of the experiment, costing between $1 and $2 million, which would increase its efficiency by a factor of three or four.

Atomic Topography

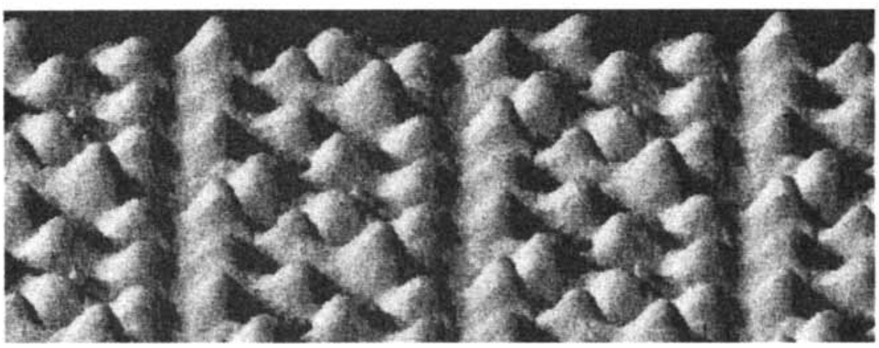
An exotic quantum-mechanical property of elementary particles has been exploited to build a new kind of microscope so powerful that it can form images of single atoms on a surface. The microscope gives unprecedented resolution in the vertical dimension and simultaneously provides a detailed picture of the relative positions of the atoms. The first such microscope was built in 1981 by investigators at the IBM Zurich Research Laboratory; similar instruments are now being tested in several other laboratories. In an image made by the Zurich laboratory [*this page*] the outer electronic shells of the surface atoms appear as rounded hills set against the valleys that form the spaces between the atoms.

The fundamental principle of the microscope is a quantum effect called tunneling. In quantum mechanics the uncertainty principle places an upper bound on the precision with which position can be known. For example, there is a finite probability that an electron nominally situated at the tip of a needle can actually be found anywhere in the vicinity of the tip. Hence the position of the electron is effectively smeared out. The probability of finding the electron near the tip falls rapidly, and in fact exponentially, with the distance from the tip.

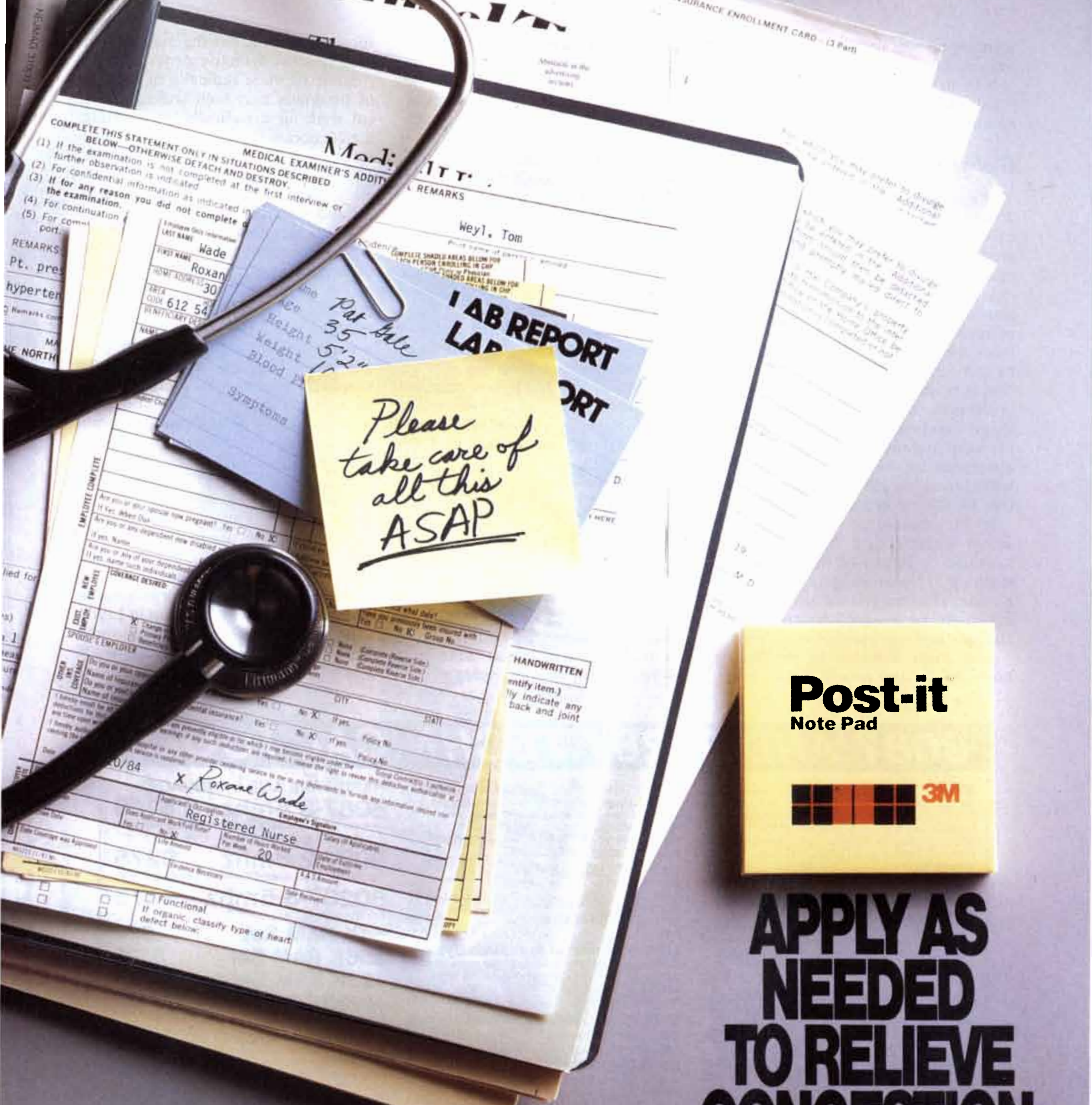
If the needle is brought close to a surface, there is a finite probability an electron once found on the needle will be found at the surface. The electron is said to have tunneled to the surface through a barrier that in classical mechanics would prevent it from making the journey. Since the tunneling probability falls exponentially with distance, the current of electrons tunneling from the needle to the surface depends very sensitively on the distance between the tip of the needle and the surface. If the distance is changed by the diameter of a single atom, the tunneling current changes by a factor of 1,000.

In practical tunneling microscopy a needle scans many times across a surface. The height of the needle above the surface atoms is kept constant by a feedback mechanism that is set to maintain a constant tunneling current. The position of the needle, which is monitored electronically, gives a record from which the topography of the surface can be reconstructed.

There is much to be gained from a better knowledge of atomic surface structure. Processes such as catalysis and crystal growth depend on details of the surface structure that are often poorly understood. The surface models that underlie the design of many electronic components may be oversimplified. It may become possible to study the atomic topography of much more complex objects such as the cell. Investigators in several laboratories are now attempting to link scanning tunneling microscopy to conventional



Scanning tunneling micrograph of silicon atoms on the surface of a crystal



APPLY AS NEEDED TO RELIEVE CONGESTION.

Post-it™ Notes quickly alleviate the symptomatic overcrowding in today's medical offices and hospitals. Our repositionable adhesive means they come off as easily as they go on. And, together with the bright color, makes certain your messages get noticed.

Call 1-800-328-1684 for a free sample. Then get more from your office manager or local stationer. With Post-it Notes, you'll give day to day operations a real shot in the arm.

Commercial Tape Division/3M "Post-it" is a registered trademark of 3M.

3M

scanning electron microscopy. If the link can be made, it will be possible to "zoom in" on predetermined surface regions that are only a few atomic diameters across.

Global Thermostat

Many climatologists predict that as combustion of fossil fuels pumps carbon dioxide into the atmosphere a greenhouse effect will warm the earth's climate. The effect results from the optical properties of carbon dioxide: it is transparent to the visible wavelengths of most incoming solar radiation but traps the infrared radiation reemitted by the earth. The amount of absorption is related to carbon dioxide concentration; climatic models predict that if the current level of carbon dioxide were to double, increased infrared absorption in the lower atmosphere would raise the global mean temperature by several degrees Celsius.

In the *Journal of Geophysical Research* Richard C. J. Somerville and Lorraine A. Remer of the Scripps Institution of Oceanography qualify this picture. They suggest the global warming might be lessened by concurrent changes in the properties of clouds.

Like most theorists of the greenhouse effect, they assume that relative humidity will stay constant as the climate warms. Warm air can hold more water vapor than cool air; if the relative humidity of the atmosphere remains stable, its total content of water vapor must increase. With more water available for condensation, the liquid water content of clouds will rise, making them denser.

Denser clouds will reflect a larger proportion of incoming solar radiation; the reduction in the energy reaching the surface will counteract the greenhouse effect. The increase in cloud density that accompanies a rise in temperature functions in effect as a climatic thermostat. Using a model that takes cloud feedback into account, Somerville and Remer calculate that doubled carbon dioxide would raise the surface temperature by only about half the amount predicted by earlier models.

Healthy Profits?

Manmade replacements for both the human heart and the human ear are currently in clinical trials. Systems intended to restore the senses of vision and touch and to provide controlled activation of paralyzed muscles are at an earlier stage of research. Advances in molecular biology could soon make possible the replacement of individual defective genes.

Some of these radically new forms of treatment may well undergo clinical trials in a radically new setting: the corporate, for-profit medical institution. According to some observers, clinical trials on human subjects in such a setting raise troublesome ethical and scientific issues.

The critics contend that profit-making corporations face strong conflicts of interest in carrying out clinical trials. A for-profit medical institution may need to promote its services or establish itself as a major purveyor of a particular treatment, the critics argue. Such business needs can lead a corporation to exploit a clinical trial as advertising, critics say. The same needs can cause a corporation to establish a more ambitious program of clinical trials than is scientifically justified, they add.

The controversy over whether clinical trials should be done at for-profit institutions was sharpened last summer when William C. DeVries moved from the University of Utah to the Humana Heart Institute International at Humana Hospital–Audubon in Louisville, Ky. DeVries is the surgeon who implanted artificial hearts in Barney B. Clark (at Utah) and William J. Schroeder (at Humana). The University of Utah is a nonprofit state institution. Humana, Inc., is a profit-making corporation that includes 89 hospitals; in 1984 Humana had gross revenues of $2.6 billion.

Humana's participation in the implantation program is a "pure marketing strategy," according to John J. Paris, associate professor of ethics at Holy Cross College. The corporation intends to make itself known as the main purveyor of artificial-heart implants, Paris said. He adds that Humana would then profit if the device is approved for commercial use.

Arnold S. Relman, editor of *The New England Journal of Medicine,* offers another critical perspective. Relman describes the implantation program at Humana as "an exploitation for commercial purposes of an experimental device that has been brought to this point by taxpayers' money in not-for-profit institutions." (Since 1964 the National Heart, Lung, and Blood Institute has provided $180 million for the development of the artificial heart and related devices.)

Humana's ambitious implantation research program may not be justified in clinical terms, according to Relman. He argues that current artificial hearts are ultimately impractical because they require the patient to be tethered to a compressor. Until a self-contained artificial heart is constructed only a few implantations should be done, mainly for the purposes of basic research, Relman says.

David Jones, president of Humana, has agreed, however, that Humana will finance 100 artificial-heart implantations. So far six procedures, including the one on Schroeder, have been approved by the Humana-Audubon institutional review board. Under the guidelines of the U.S. Food and Drug Administration, which regulates trials of experimental medical devices, the review board is responsible for evaluating the ethical aspects of research and ensuring that informed consent is obtained. The FDA has also approved the six implantations. The regulatory agency is satisfied that the Humana-Audubon board did its work conscientiously, FDA officials say.

Robert B. Irvine, a spokesman for Humana, disagrees strongly with Paris and Relman. Humana decided to fund the implantation program in the interests of scientific knowledge and not out of commercial self-interest, Irvine says. Humana did not seek publicity for the procedure on Schroeder but simply tried to accommodate the very high level of existing journalistic interest, he adds. "Rather than public relations," Irvine comments, "I'd call it media anticipation."

Irvine notes that in agreeing to provide funds for 100 implants Jones stipulated that the trials would end if at any time they ceased to contribute to the advancement of medical knowledge. Irvine also points out that the 100 implantations funded by Humana are not limited to currently available artificial hearts. If better and more practical devices become available in the future, they will be employed, he says.

The issues in the current controversy will continue to be publicly debated, but they are not likely to be resolved through Government action. The reason is that FDA regulations do not distinguish between for-profit and not-for-profit institutions, agency officials say. According to William G. Letzing, chief of prosthetic devices in the cardiac-device unit of the Office of Device Evaluation: "Nothing in the FDA regulations indicates that a for-profit institution is to be considered any lower than a university hospital. Our mission in a nutshell is to make sure the device is safe and effective, that the patient is well aware of what he's getting into and that the institutional review board is well informed."

Brownian Motion

Observing the random course of a particle suspended in a fluid led to the first accurate measurement of the mass of the atom. Brownian motion now serves as a mathematical model for random processes

by Bernard H. Lavenda

It sometimes happens that a drop of water is trapped in a chunk of igneous rock as the rock cools from its melt. In the early 19th century the Scottish botanist Robert Brown discovered such a drop in a piece of quartz. The water, Brown reasoned, must have been inaccessible for many millions of years to spores or pollen carried by the wind and rain. He focused a microscope on the drop of water. Suspended in the water were scores of tiny particles, ceaselessly oscillating with a completely irregular motion. The motion was familiar to Brown: he had previously happened to observe such oscillations during his studies of pollen grains in water. The new experiment, however, ruled out the explanation he had put forward earlier, namely that "vitality is retained by [the 'molecules' of a plant] long after the [plant's] death." Brown rightly concluded that the agitation of the particles trapped inside the quartz must be a physical phenomenon rather than a biological one, but he could not be more specific.

The correct explanation for the so-called Brownian motion is now well known. A grain of pollen or dust suspended in a fluid is continually bombarded by the molecules that make up the fluid. A single molecule hardly ever has enough momentum for its effect on the suspended particle to become visible under a microscope. Nevertheless, when many molecules collide with the particle from the same direction at the same time, they can noticeably deflect the particle.

Brownian motion is consequently a doubly random effect: the path of the suspended particle is randomized by random fluctuations in the velocities of nearby molecules. Moreover, because the microscope is essentially a filter that portrays only the effects of relatively large fluctuations in the local molecular environment, the observed motion only begins to suggest the complexity of the true path. If the resolving power of the microscope were increased by factors of, say, 10, 100 and 1,000, the effects of bombardment by progressively smaller groups of molecules could be detected. For each increase in the magnification, parts of the trajectory of the particle that initially appeared to be straight lines would be seen to have a jagged and irregular structure. The path of a particle in Brownian motion was one of the first natural phenomena recognized as being effectively self-similar at every magnification, a key property of geometric objects Benoit B. Mandelbrot of the IBM Thomas J. Watson Research Center has named fractals.

Probabilistic Phenomena

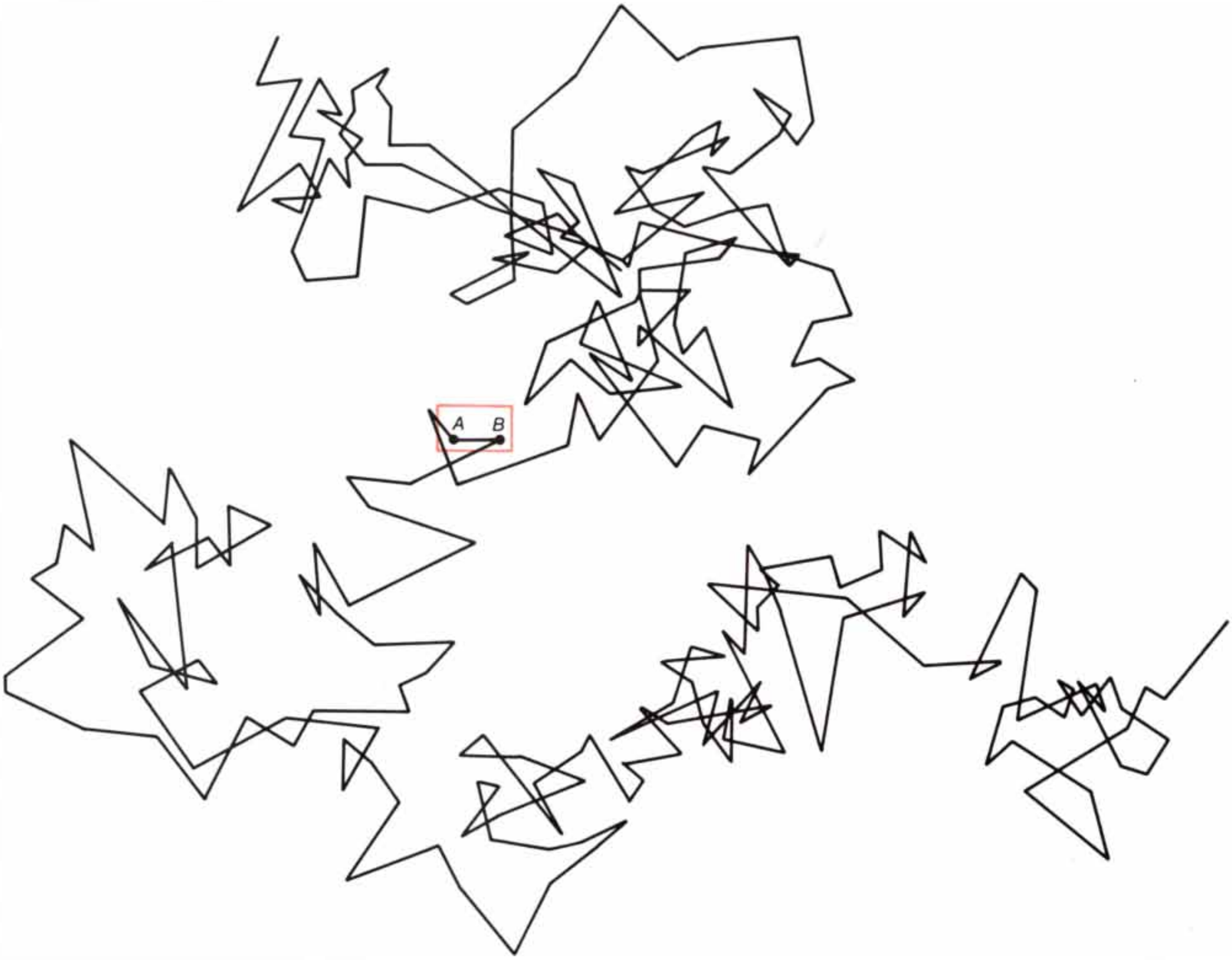
Since the turn of the century the study of Brownian motion has had far-reaching consequences for physics, chemistry and mathematics. In the hands of Albert Einstein it was turned into a conclusive, observational method for confirming the atomic theory of matter. Moreover, Einstein showed how the measurement of certain properties of particles in Brownian motion could determine several important physical constants: the masses of atoms and molecules and the magnitude of Avogadro's number, which is the number of elementary particles in a mole, the standard chemical unit of any substance. Brownian motion also contributed to a deepened theoretical understanding of thermodynamic principles, which had previously been formulated on the basis of what turned out to be oversimplified empirical generalizations.

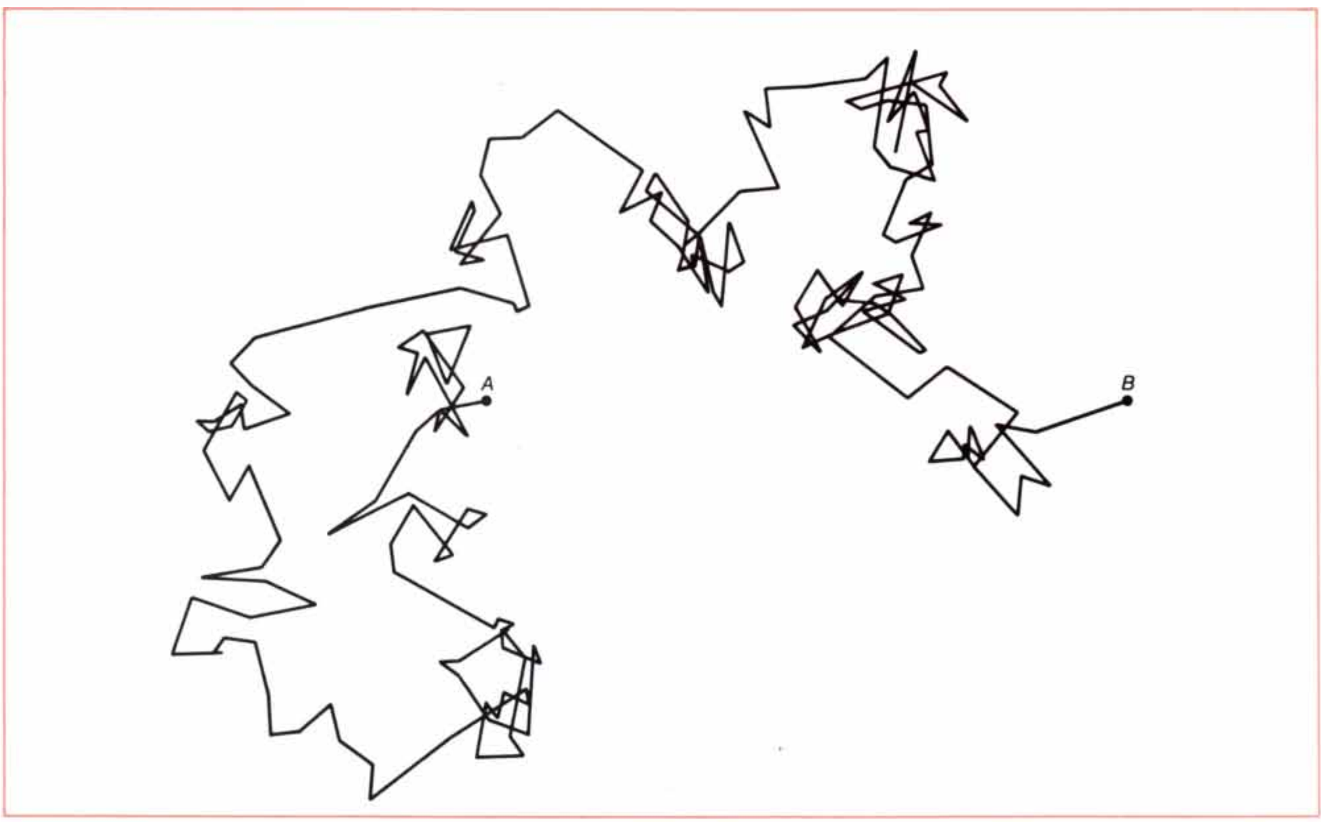
More recently the study of Brownian motion has led to the invention of important mathematical techniques for the general investigation of probabilistic processes. The techniques have been applied in the control of electromagnetic "noise," and they have contributed to the understanding of the dynamics of star clustering, the evolution of ecological systems and the behavior of stock prices.

Surprisingly, Brownian motion did not attract much attention in the 19th century. Scientists of the day dismissed the phenomenon as an effect of local thermal currents brought about by minute differences in the temperature of the fluid. If the motion were caused by such currents, neighboring particles would be dragged along by the same local current, and so one would expect such particles to move in roughly the same direction. Under the microscope this expectation is completely contradicted. The suspended particles have no effect on one another, even when they are separated by distances smaller than their diameters.

By the turn of the century several experimental findings had begun to hint at the molecular origin of Brownian motion. For example, it was known that the smaller the size of the particle, the more rapid its Brownian motion. An increase in the temperature of the fluid also seemed to cause more agitated Brownian motion. Such effects were recognized as being consistent with the

RANDOM, BROWNIAN MOTION of a microscopic particle suspended in water is shown in the upper illustration as it was plotted in 1912 by the French physicist Jean Baptiste Perrin. Perrin's diagram records the position of the particle every 30 seconds; such diagrams, as he noted, give "only a very meager idea of the extraordinary discontinuity of the actual trajectory." If a small part of the trajectory is magnified and the position of the particle is marked, say, 100 times more frequently, the complexity of the original trajectory is reproduced (*lower illustration*). The lower diagram is based on a simulation done by the author.


A
B


A
B

kinetic theory of gases soon after its development by James Clerk Maxwell and Ludwig Boltzmann in the 1870's. Nevertheless, it was not until 1905 that Einstein stated the first quantitative implications of the kinetic theory for Brownian motion.

The Kinetic Theory

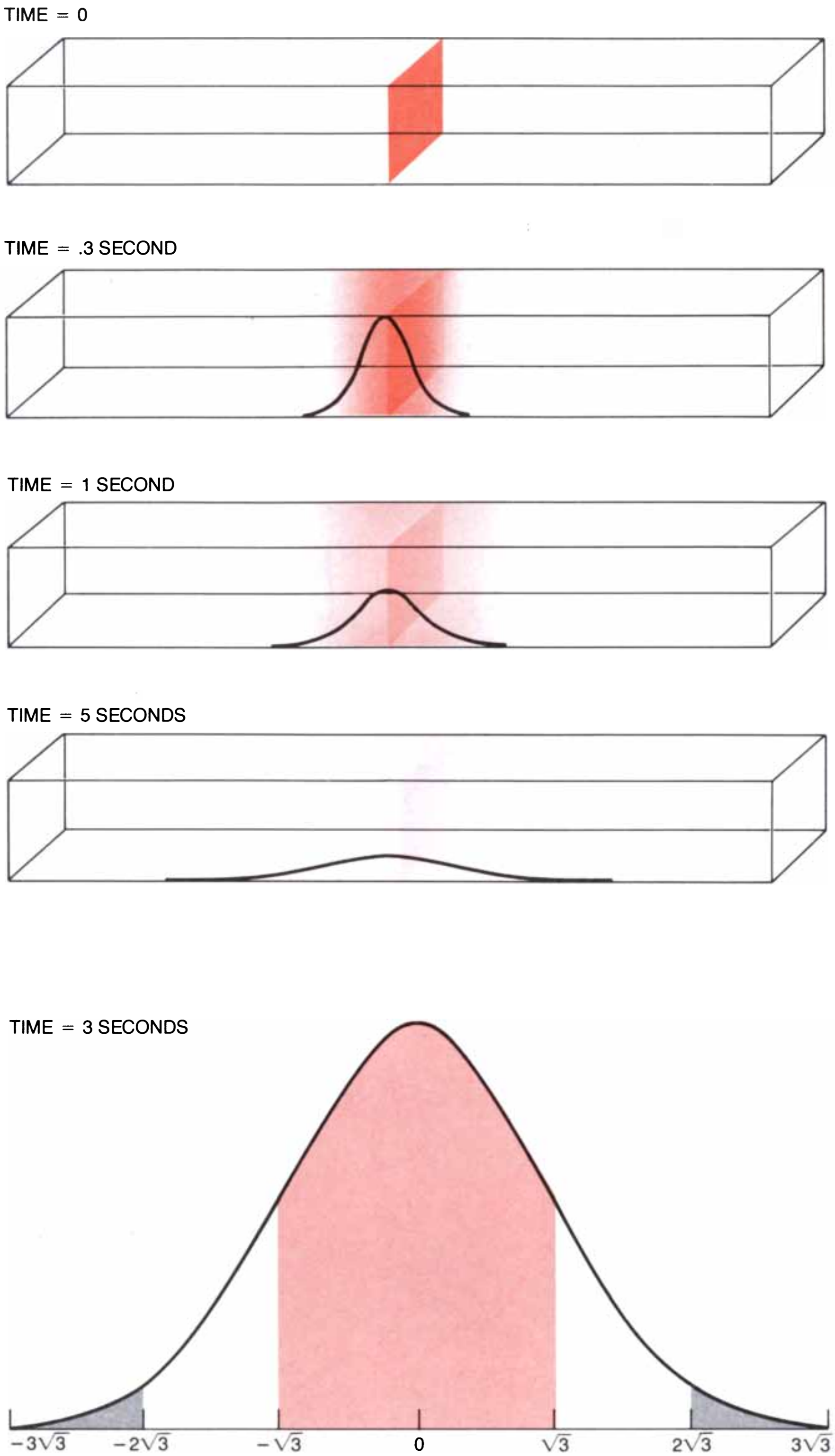
The kinetic theory of gases was the first successful attempt to explain the well-known properties of a gas as the macroscopic effects of atoms in motion. For example, it had been known since Robert Boyle's investigations in the 17th century that there is an inverse relation between the pressure and the volume of a gas. If the volume is decreased but the temperature is not changed, the pressure increases by a proportional amount. If the volume is increased, the pressure decreases. According to the kinetic theory, pressure is a result of constant bombardment by the particles against the walls of their container. The pressure increases with decreasing volume because the rate of bombardment of the particles is greater for a small volume than it is for a large one.

Similarly, there is a direct relation between pressure and temperature. If the temperature of a gas is increased but its volume is not changed, the pressure increases proportionally. If the temperature is decreased, the pressure decreases. Temperature is understood by the kinetic theory as being a measure of the average kinetic energy of the particles. A higher temperature is equivalent to an increase in the energy of the average bombardment, and so it increases the pressure of the gas.

For a so-called ideal gas these two relations are summarized by a simple law. The law states that for one mole of a gas the product of the pressure and the volume of the gas divided by its absolute temperature is equal to a constant. The constant is called the universal gas constant and is designated by the letter R; it is equal to 1.99 calories per mole per degree Celsius.

The most important conceptual advance of the kinetic theory was that it gave up any pretext of providing a detailed description of the motions of individual particles. Instead it offered a statistical account of the motion, which was made plausible by the fact that a system made up of many particles is highly unlikely to deviate significantly from its average behavior. Accordingly, the kinetic theory is often called statistical mechanics.

In retrospect it may seem intuitively obvious that a dust particle or a pollen grain embedded in the turbulent atomic environment of a gas or a liq-



DIFFUSION of Brownian particles (*color*) through a transparent liquid or gas is shown at successive times. The particles begin diffusing from a permeable membrane inserted in the middle of the box. The bell-shaped curve on each box is a graph of the relative density of the particles for each point along the horizontal dimension of the box. The root-mean-square displacement, or most probable displacement, of a particle after a given time has elapsed is proportional to the square root of the time. The bell-shaped curve at the bottom of the illustration shows how the root-mean-square displacement is related to the probability that a given particle will occupy some region of the box after a given time. For example, if the root-mean-square displacement after three seconds is $\sqrt{3}$ centimeters, the probability that a particle is found within $\sqrt{3}$ centimeters of the central membrane is equal to the area of the colored region under the curve, which is about .68. The probability that the particle is found farther than $2\sqrt{3}$ centimeters from the central membrane is equal to the combined area of the two gray regions, which is less than 5 percent of the area under the entire curve.

THE McGRAW-HILL ENCYCLOPEDIA OF ELECTRONICS AND COMPUTERS

YOURS FOR ONLY $2.95!

WHEN YOU JOIN THE LIBRARY OF COMPUTER AND INFORMATION SCIENCES.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

Everything you want to know about computers is clearly, simply explained in the *McGraw-Hill Encyclopedia of Electronics and Computers.* With 964 double-columned pages, 477 articles by 272 contributors, and 1266 illustrations, tables, and graphs that clearly illustrate the how, the why, and the what.

Written by the most eminent authorities of the academic and industrial world, with substantial bibliographies for further study, the Encyclopedia covers such topics as: algorithms, analog computers, circuit theory, computer graphics, computer storage technology, data communications, disk recording, ferroelectronics, graph theory, information theory, lasers, optics, switching systems, and video disk recording. And that's just the beginning.

Join *The Library of Computer and Information Sciences* to keep up-to-date on the developments, theories, and applications important to your job.

The Library of Computer and Information Sciences offers an unmatched data bank for any computer professional. As the oldest, largest, and most respected club of its kind, it brings you the finest publications from the most authoritative writers in the field. *All at significant savings!*

As a member, you won't have to spend time sorting through hundreds of books to find the ones worth reading. The Club's editors do it for you. Reviewing the hundreds of books submitted monthly, they select only those books that offer the critical, hands-on information you need. Books that give you an in-depth knowledge on a variety of exciting subjects—from systems design to applications, from software engineering to personal computing. You'll regularly be offered critically important books you might otherwise have missed.

Take advantage of our 10-day, no-risk examination period.

Join the Club risk-free, simply by sending for the *McGraw-Hill Encyclopedia of Electronics and Computers* now. When it arrives, examine it for 10 days. If you are not completely satisfied, return the book and owe nothing. Otherwise, we'll bill you just $2.95.

Simply complete and mail the card attached.

4 Good Reasons to Join

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting the McGRAW-HILL ENCYCLOPEDIA OF ELECTRONICS AND COMPUTERS for $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Library of Computer and Information Sciences News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed, please write to: **The Library of Computer and Information Sciences, Dept. 7-CV6-61613, Riverside, N.J. 08075,** to obtain membership information and an application.

Scientific American 2/85

THE WORLD NEEDS A PLACE LIKE NO PLACE

A manned space station will offer governments, industries and people opportunities that cannot be matched on Earth.

It will make possible technology advances held back by Earth's limitations. The space environment offers a weightless, vibrationless, moistureless, soundless vacuum in and with which to work; an obstruction-free perspective for scanning our planet and a host of other out-of-this-world advantages.

A manned space station makes good business sense. A space station can revolutionize medicine, metallurgy, communications, energy generation, meteorology and dozens of other sciences and technologies—ones that exist now, and new ones a space station can help create.

A manned space station will help promote understanding. It will offer the people of the world an opportunity to unite in peaceful enterprise.

And a manned space station will help America maintain its momentum in space science and exploration.

Working to make the space station a reality.

For almost three decades, one company has been at work to make a manned space station operational. The name of that company is McDonnell Douglas.

Since 1959, space station studies have continued at McDonnell Douglas Astronautics Company without interruption—some commissioned by the government, others as our own investment. The thrust was always to set the pace. We've done analytical tool development. Design and test activities. Metallurgical studies. Wet workshop studies. Space cabin simulator operations. Thermal control development.

McDonnell Douglas' unwavering support for a human role in space is exhibited by its space life sciences department, maintained continuously for two and a half decades. We built the Mercury and Gemini spacecraft for NASA; provided launch stages for Apollo moon missions; and built and flew unmanned launch systems and spacecraft, learning to integrate them

ON EARTH—NASA SPACE STATION.

with payloads and systems built by others.

Skylab was a forerunner.

McDonnell Douglas was NASA's contractor for Skylab, the first device that the world could call a space station. It was a gigantic system for its time. In it, American government and industry first explored the technologies and experimental projects that will find fulfillment in a space station of the 1990s.

But we need not dwell on laurels. Management, system integration and technical prowess gained in early programs continue: Payload integration on the space shuttle; shuttle simulation programs, astronaut training, and mission planning for NASA. In aerospace programs throughout McDonnell Douglas, we draw on the skills of other major companies across America. We are NASA's integration contractor for the multi-nation European Spacelab, and we celebrate ongoing successes with international industry partners on a dozen different space, aviation and missile programs.

An experienced team.

We have teamed with Honeywell, IBM and RCA for a major segment of a NASA contract for definition and preliminary design of the manned space station. And already we're developing the pharmaceutical technologies and the production factories that will find their full potential in the space station environment.

At McDonnell Douglas, our capabilities are focused on making a space station a reality by the next decade—and we offer exciting opportunities to companies and the men and women in them who want to help.

Because the world needs a place like no place on Earth.

MCDONNELL DOUGLAS

SCIENCE/SCOPE

Satellite images have led to the discovery of a large Mayan city hidden for centuries by the jungles of Mexico's Yucatan peninsula. Scientists uncovered over 100 possible ancient sites by studying false-color images provided by Landsat earth resources satellites. Jungles typically are made to appear bright red in these pictures. The ruins appeared pink or light red, because foliage had not grown as densely over the sites as it had in the neighboring jungle. The Landsat images also revealed the existence of vast farmed fields that may prompt new theories of how the ancient civilization was able to feed itself. The imaging instruments aboard the Landsat satellites, called multispectral scanners, were built by the Santa Barbara Research Center, a Hughes Aircraft Company subsidiary.

Mexico is a step closer to its own domestic communications system after taking delivery of a satellite control facility. The system will consist of two Hughes HS 376 satellites and the control facility, which will monitor the spacecraft as they are maneuvered into geosynchronous orbit 22,300 miles above the equator. The facility is being installed in Iztapalapa, a suburb of Mexico City. Upon completion, a Hughes support team will remain for 18 months to train Mexican engineers and technicians. The program, called Morelos, is named after a Mexican patriot of the early 19th century. The satellites are scheduled for launch from the space shuttle in May and November.

A high-frequency tactical Manpack radio used by all branches of the U.S. military is proving its reliability in the field, operating on average more than 4,000 hours between failures. The AN/PRC-104 high-frequency radio makes extensive use of large-scale integrated circuits, conservatively rated components, and proven military equipment packaging techniques. Should it need repairing or maintenance in the field, an operator can replace any of the three basic subsystems in seconds. The average repair time in the field is less than 20 minutes. In production at Hughes for the U.S. Army, Marines, and Air Force, the radio is available for international needs.

Improvements to a "super cooler" used with infrared sensors in space will extend the life and boost the efficiency of the device. The cooler, vital to defense applications and geological surveys, is a Vuilleumier cycle cryogenic refrigerator. It is designed to chill sensors near absolute zero to increase their sensitivity to thermal radiation. These coolers are ideal for use in space because the low internal forces required by this kind of cooling cycle cause little wear on bearings and seals. Hughes is working under a U.S. Air Force contract to extend the unattended operating life of the cooler beyond five years. The cooler will use less power, so smaller and fewer batteries are needed to power the device during eclipse periods—a saving of hundreds of pounds.

The Hughes Tucson facility, located in picturesque Southern Arizona, is a large, modern manufacturing complex with capabilities for producing advanced missile systems developed by Hughes. We have openings for experienced and graduating engineers to work on such advanced systems as the electro-optical Maverick, radar-guided Phoenix, TOW, and AMRAAM, the Advanced Medium-Range Air-to-Air Missile. Please send your resume to Professional Employment, Dept. S2, Hughes Aircraft Company, P.O. Box 11337, Tucson, AZ 85734. Equal opportunity employer. U.S. citizenship required.

For more information write to: P.O. Box 11205, Dept. 65-10, Marina del Rey, CA 90295

HUGHES
AIRCRAFT COMPANY

uid must undergo Brownian motion. To appreciate Einstein's contribution, however, one must be aware that the physical reality of the atom and its constituents now taken thoroughly for granted had only provisional status 80 years ago. The German physicist Wilhelm Ostwald still regarded atoms as merely "a hypothetical conception that affords a very convenient picture" of matter. Ernst Mach maintained that all theoretical entities, including atoms and molecules, must be treated as convenient fictions.

Einstein's Theory

Einstein was a realist in the matter. He wrote in 1905 that the original goal of his investigations of atomic theory "was to find facts that would guarantee as much as possible the existence of atoms of definite size. In the midst of this," Einstein continued, "I discovered that, according to atomistic theory [that is, the kinetic theory], there would have to be a movement of suspended particles open to observation, without knowing the observations concerning Brownian motion were already long familiar." Thus Einstein was the first with the physical insight to recognize that atoms would reveal their existence in the motions of particles suspended in a fluid. Although he was ignorant of Brownian motion, he argued that the detection of such particles would constitute a strong confirmation of the kinetic theory. Among the most startling conclusions of his work is an equation that made it possible for the first time to accurately measure the mass of the atom.

There are two main parts to Einstein's investigation of the atomic theory of Brownian motion. The first part is mathematical; an equation is derived that describes the diffusion of a suspended Brownian particle through a fluid medium. The second part is a physical argument, and it relates the measurable rate of the diffusion of the particle to other physical quantities, such as Avogadro's number and the universal gas constant.

To express the diffusion of a particle in the mathematical language of classical mechanics one generally must know two things: the initial velocity of the particle and the magnitude and direction of the impulses the particle receives in a given time. Because a Brownian particle undergoes about 10^{21} collisions per second, any effect its initial velocity might have on its subsequent motion is overwhelmed by the effects of molecular collisions in a vanishingly short time. It is also hopeless to describe the impulses individually. Einstein therefore abandoned a straightforward mechanical description of the diffusion of a Brownian particle. Instead he introduced a probabilistic account.

In order to derive Einstein's result it is useful to imagine a small volume having any arbitrary shape within the region accessible to the diffusing particles. The rate at which the concentration of diffusing particles changes with time is equal to the flow into the region less the flow out. Each flow depends on the flux of particles, or flow per unit area. The flux of particles in a fluid from point to point is directly proportional to the difference in the concentrations of the particles at the two points. The coefficient of proportionality is called the diffusion coefficient, or *D,* and its value must be determined experimentally. The relation between flux and the rate of change in the concentration is known as Fick's law, after Adolph Fick.

The mathematical formulation of this physical state of affairs leads to a differential equation called the diffusion equation. The equation can be solved if the initial position of the diffusing substance and the boundaries of the space accessible to the substance are specified. The solution is a mathematical expression that gives the concentration of the diffusing substance at every point in the space for every time. For example, if the diffusing substance is initially concentrated on the surface of a permeable membrane that divides a box in two at its center, the solution of the diffusion equation is a family of bell-shaped curves. The center of each curve coincides with the center of the box, and as time passes the curve becomes broader and flatter [*see illustration on page 72*].

Particle Displacement

There is another way to interpret each bell-shaped curve. Every point on the curve can be considered the probability density for the diffusion of a single Brownian particle from the central membrane in the box. The choice of the metaphorical term probability density is appropriate because the density of an ordinary substance multiplied by its volume is the amount of the substance. Similarly, the probability density multiplied by an appropriate measure gives rise to a probability. For the bell-shaped curve the appropriate measure is length instead of volume: it is the distance between two points along the horizontal axis of the graph. The product of that distance and the average height of the bell-shaped curve in the interval between the two points is a probability.

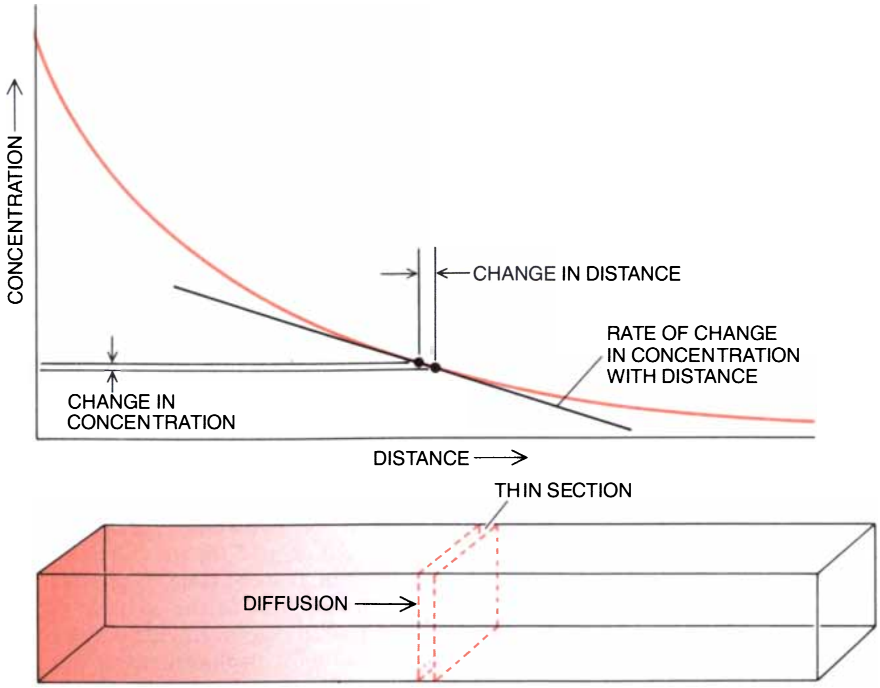
The probability that the Brownian particle will be found in a given region of the box at a given time is the area under one of the bell-shaped curves between two vertical lines. Each vertical line passes through one of the two points on the horizontal axis that correspond to the boundaries of the given region in the box. According to this interpretation, the solution of the diffusion equation is not an expression that gives the distribution of particle concentrations; instead the solution is a probability distribution.

A convenient probabilistic measure of the displacement of a Brownian particle is the root-mean-square displacement. The easiest way to understand the root-mean-square displacement is to consider the diffusion of a large number of Brownian particles. The displacements of all the particles are measured at some time t and squared; the square root of the average of all the squared displacements is the root-mean-square displacement at time t. The probability that a Brownian particle diffuses at most as far as the root-mean-square displacement away from the central membrane in the box is .68; the probability that it travels more than twice as far as the root-mean-square displacement is less than .05.

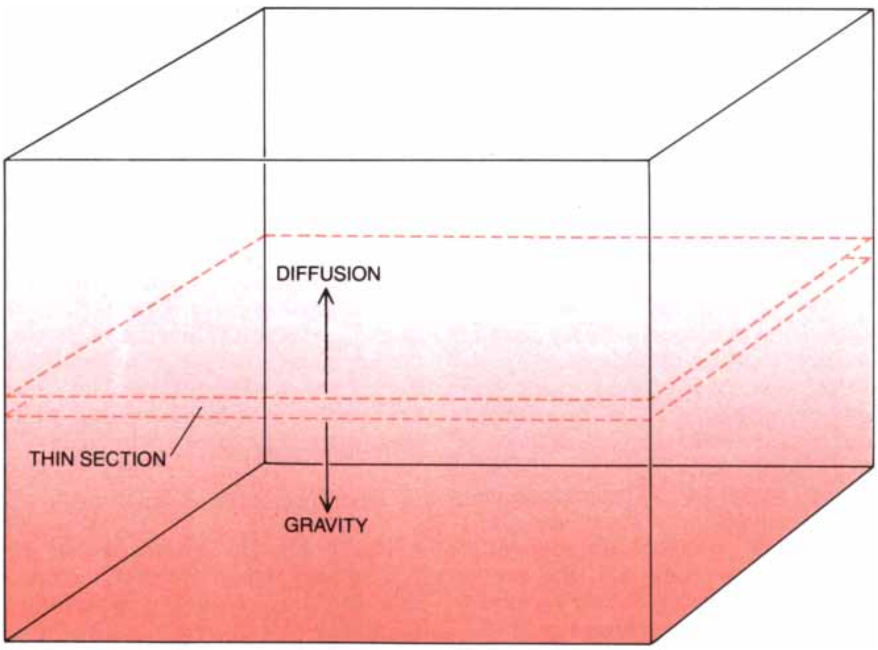
The root-mean-square displacement of a Brownian particle diffusing away from the membrane in the box is $\sqrt{2Dt}$, where D is the diffusion coefficient and t is the time. Thus if a particle diffuses on the average one centimeter in one second, it will require four seconds to diffuse two centimeters and nine seconds to diffuse three centimeters. Other initial conditions on the differential equation for diffusion give rise to similar solutions; in fact, Einstein was able to predict that the radial displacement of a particle diffusing in any direction away from a central point "is not proportional to time but proportional to the square root of time." The result follows, he asserted, "from the fact that the paths described during two consecutive time intervals are not always to be added, but just as frequently have to be subtracted."

The Measurement of Diffusion

This prediction made possible the first serious test of Einstein's expression for the probability distribution for the displacements of a Brownian particle. The French physicist Jean Baptiste Perrin and his students followed the movement of a nearly spherical Brownian particle and recorded its position at equal intervals of time. After repeating the experiment many times, they plotted the mean-square displacement against time. The graph they obtained was a straight line; the



NONUNIFORM CONCENTRATION of Brownian particles in a box (*color*) gives rise to a force of diffusion that drives the particles in the direction of lower concentration. In the upper illustration the concentration of the particles at a given time is plotted against distance. The rate of change in concentration from point to point along the box is given by the slope of the line connecting the two points on the curve that correspond to the two points on the box. The average flux of particles across any thin section of the box is proportional to the rate at which the concentration of the particles changes with distance; hence the flux is proportional to the slope of the line that connects the two points on the concentration curve corresponding to the boundaries of the thin section. Flux is also equal to the average concentration of the moving particles across the thin section multiplied by their average velocity.



EQUILIBRIUM DISTRIBUTION of Brownian particles in a gravitational field arises from a dynamic balance between gravity and the force of diffusion driving the particles from the higher concentration at the bottom of the column to the lower concentration at the top. Across any given slice of the column the average velocity of the particles moving downward under gravity is equal to the average velocity of the particles diffusing upward. The upward velocity is proportional to the rate of change in the concentration of the particles with height (*see top illustration on this page*). If one assumes that the particles form a so-called ideal gas, it is mathematically straightforward to derive an algebraic expression for the weight of an atom or a molecule in terms of experimentally measurable quantities.

slope of the line was *D*, the diffusion coefficient.

The experimentally determined value of *D* is the one new quantity introduced in Einstein's theory that makes it possible to measure the size of the atom. Imagine that microscopic granules or Brownian particles have been suspended in a vertical column of still air [*see bottom illustration at left*]. Because the granules are subject to the force of gravity, they tend to fall to the bottom of the column. As their concentration increases near the bottom of the column, however, the concentration difference between the top and the bottom tends to drive the granules toward the top of the column, where the concentration of granules is relatively low. When equilibrium has been established, the distribution of the granules reflects a balance between the downward pull of gravity and the upward push of diffusion.

The viscosity of the air prevents the granules from being continuously accelerated downward by gravity. Instead they reach a certain terminal velocity and then drift downward at a constant speed, which is equal to the force of gravity on each particle divided by the viscosity. Because viscosity changes with the increase in air density, the terminal velocity varies with height, but that complication can be eliminated by considering only the velocities of particles at some arbitrary, given height. At equilibrium the number of granules drifting down past the given height on the column must be matched by the number of granules drifting up. Consequently at the given height the average velocity of a granule moving up the column must be equal to the terminal velocity of the downward-moving granules.

The flux of the diffusing granules is equal to their average velocity divided by the volume of the small cloud of granules that crosses the given height on the column of air in a short time. Flux can therefore be expressed as the product of the velocity of the diffusing granules and their concentration *C* at the given height. Since the flux is also equal to the product of *D*, the diffusion coefficient, and the rate at which the concentration of the granules is changing with height, the average velocity of a granule diffusing up the column is equal to D/C multiplied by the rate of change in concentration with height.

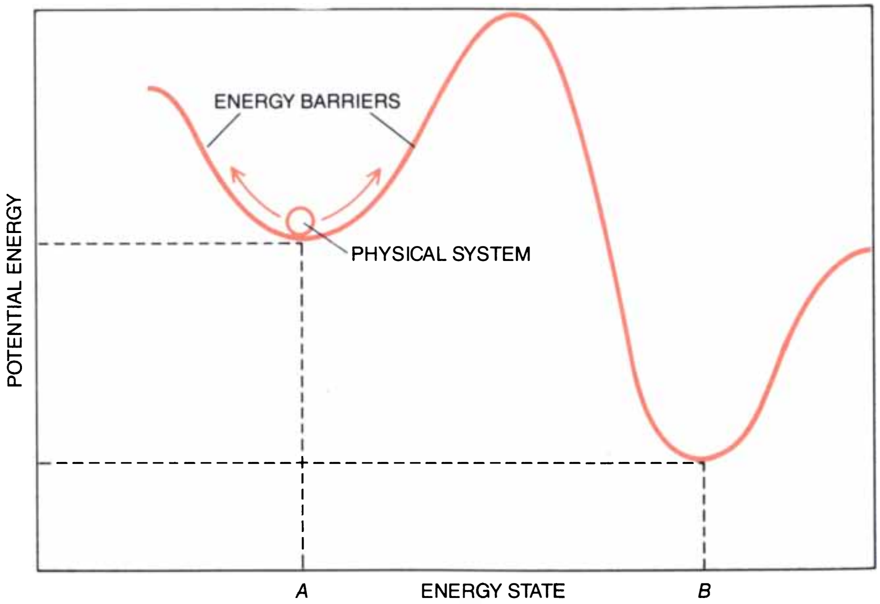
The Mass of the Atom

The rate of change in concentration varies directly with the rate at which the pressure of the granules is changing with height, and the constant of proportionality is derived from the law of ideal gases: it is equal to N_0/RT. where N_0 is Avogadro's number, R is the universal gas constant and T is the absolute temperature of the gas. The appearance of Avogadro's number at this stage in the argument follows purely from the law of ideal gases, but when Einstein gave his argument, it was still an imprecisely measured quantity. Its introduction in this context made it possible to link two properties of fluids that had previously seemed unrelated: the motion of a particle through a viscous medium and the so-called osmotic pressure that is exerted by a dissolved substance confined by an outside force to a small region within a fluid.

The rate at which the pressure is changing with height at any given height is the product of the concentration of granules in a thin layer just above the given height and the force of gravity on each granule. By setting the velocity of a diffusing granule equal to the terminal velocity of that granule moving under gravity, a straightforward algebraic manipulation gives Einstein's expression for the diffusion coefficient: $D = RT/fN_0$, where f is the viscosity of the air. From this equation the value of Avogadro's number is RT/Df, or the product of the universal gas constant and the absolute temperature, divided by the diffusion coefficient and the viscosity.

A numerical determination of Avogadro's number is tantamount to a measurement of atomic dimensions. The first person to exploit Einstein's result for this purpose was Perrin, and so Perrin can be said to be the first to weigh the atom. The viscosity f can be measured in many ways, say by measuring the mean velocity of a particle falling through a fluid. The universal gas constant R is determined by measuring the temperature and the pressure of a known quantity of gas confined to a known volume. Since Perrin had already measured D and since T is measured by a thermometer, the value of Avogadro's number can be calculated. It is roughly 6×10^{23}.

In view of the number of assumptions that were made by Einstein it is remarkable that Perrin's value for Avogadro's number is within 19 percent of its currently accepted value. The magnitude of Avogadro's number leads directly to the mass of the atom and the molecule. By definition the mass of an elementary particle is the weight of one mole of the substance made up of such particles divided by Avogadro's number. For example, a mole of oxygen gas weighs 16 grams. The weight of a single oxygen molecule is therefore $16 / (6 \times 10^{23})$, or 2.7×10^{-23} grams. The painstaking work that led to this calculation and similar ones was the final blow to those who remained skeptical about the atomic theory. Perrin was awarded the Nobel prize in physics in 1926.



ENERGY DIAGRAM is plotted for a physical system having two states, *A* and *B*, for which the potential energy is at a minimum with respect to closely neighboring states. In the absence of thermal fluctuations, if the system were to occupy either state *A* or state *B*, it would remain trapped there indefinitely. If every physical state is accompanied by random fluctuations, however, there is a finite probability that the system will not remain in any locally minimum state. For example, a system in state *A* may cross the energy barrier and reach state *B*, and a system in state *B* may reach state *A*. By comparing probabilities for the two transitions the relative stability of the two locally minimum states can be compared.

Thermodynamics

What I have said so far about Einstein's theory of Brownian motion does not do it full justice. Not only did it verify the physical existence of atoms but also its spectacular success established statistical mechanics as the base on which all the laws of thermodynamics had to stand or fall.

Given the size of Avogadro's number and the minuteness of the atom, it is not hard to understand why it was possible even before the development of statistical mechanics to state phenomenological, or macroscopic, thermodynamic laws that were approximately correct. According to the law of large numbers, large fluctuations, or deviations, from average behavior in a macroscopic system of 10^{23} particles must be rare. The smaller and commoner fluctuations predicted for macroscopic systems were too small for the relatively insensitive measuring devices of the 19th century. Nevertheless, the statistical view of thermodynamics implied fundamental revisions to the phenomenological thermodynamic laws.

For example, the perpetual movement of a particle undergoing Brownian motion contradicts the earliest, phenomenological version of the second law of thermodynamics. According to the phenomenological second law, the temperature of everything in some closed environment tends to shift everywhere toward the same value; when this equilibrium is reached, there is no way for the thermal energy of the environment to be transformed into useful energy, or work.

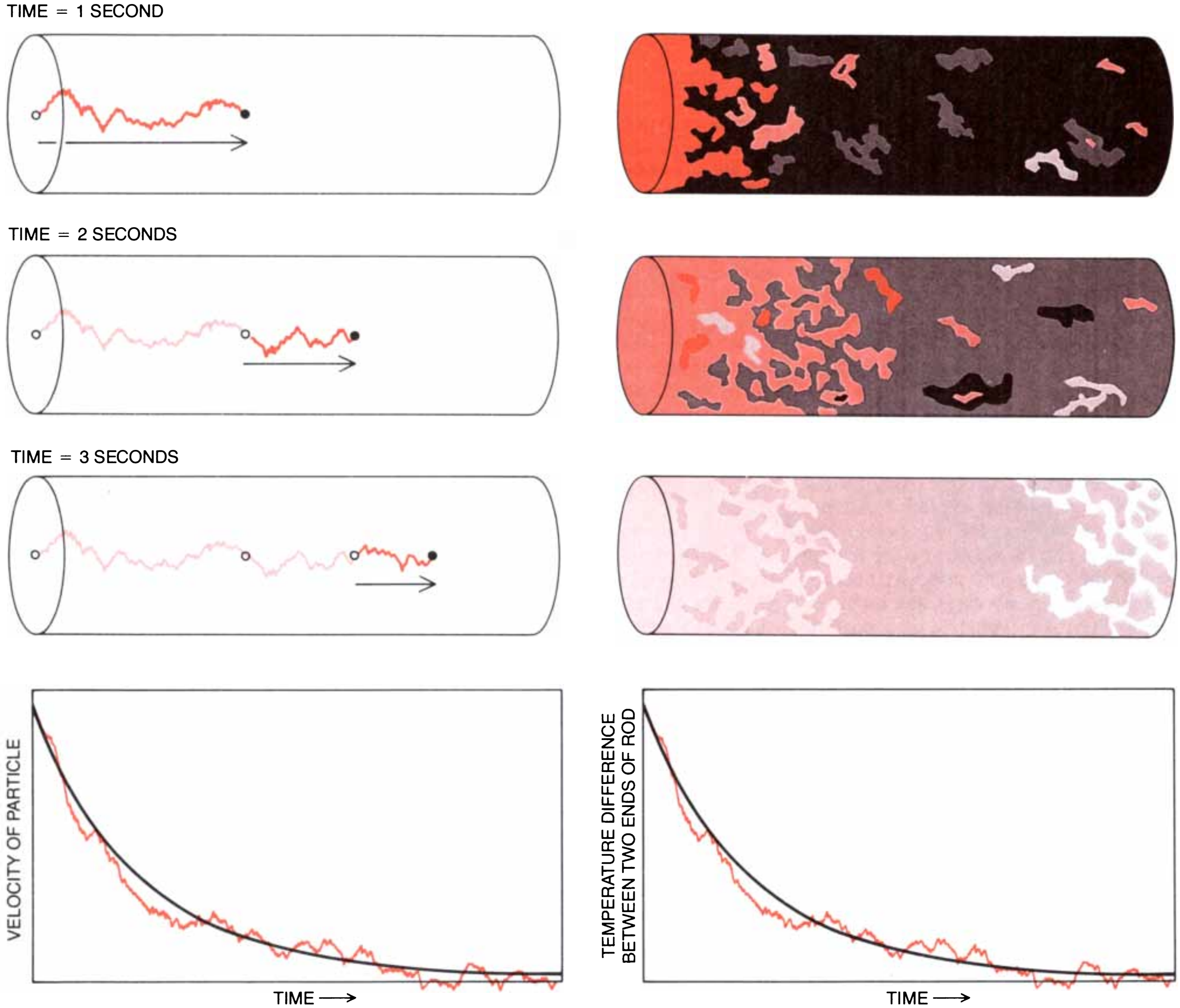
Now, the temperature of a Brownian particle suspended in water is the same as that of the water, but the kinetic energy of its constant jiggling must be derived from the kinetic energy of the water molecules. Because temperature is simply a way of expressing the translational kinetic energy of the molecules, the brief transfer of kinetic energy to the Brownian particle can be accomplished only by a local cooling of the water. Thus Brownian motion shows that the state of completely uniform temperature presupposed by the phenomenological second law is never satisfied in nature. The definition of thermodynamic equilibrium must instead take account of small but persistent, random fluctuations in the temperature of a system.

The importance of such fluctuations in physical systems in equilibrium was first recognized by Einstein in 1910. He developed a theory of fluctuations by introducing statistical concepts into thermodynamics. Phenomenological thermodynamics simply codified the repeated experimental observations that energy is always being degraded, in the sense that it becomes increasingly difficult to extract for doing useful work. In phenomenological thermodynamics the measure of the degradation of energy was called entropy, from the Greek words meaning internal change. According to the phenomenological second law, the entropy of any system always tends toward a maximum with the passage of time.

In statistical thermodynamics the definition of entropy was subtly altered. According to the atomic theory, all macrostates, or macroscopic, observable states of a system, including the equilibrium state, arise from various arrangements of atoms and molecules. These arrangements are called microstates. In any isolated experimental system every microstate consistent with a given macrostate is equally probable. It is reasonable to suppose the equilibrium state identified by the phenomenological second law is the macrostate for which there is the greatest number of microstates.

There are far more essentially indistinguishable microstates associated with a disordered macrostate than there are microstates associated with an ordered macrostate. Hence if equilibrium in statistical thermodynamics was to continue to be the state of maximum entropy, entropy had to be redefined as a measure of the disorder of a physical system instead of the degradation of its energy. The more disorder there is in the system, the more microstates there are that collapse into a single macrostate and the greater the entropy of the macrostate is.

The precise relation between the entropy of a macrostate and the number of associated microstates was formulated by Boltzmann in 1896. The entropy is proportional to the logarithm



APPROACH TO EQUILIBRIUM TEMPERATURE of a rod heated at one end (*right*) is mathematically analogous to the motion of a grain of sand moving through a liquid (*left*). The drag force on the grain caused by the viscosity of the liquid is analogous to the "force" that tends to restore the heated rod to a uniform temperature. The drag force is proportional to the velocity of the grain, and so it diminishes continuously as the grain slows down. Similarly, the force that tends to restore the equilibrium of the heated rod is proportional to the temperature difference between the two ends; as that difference diminishes, the restorative force is also reduced. Both processes can be shown graphically as smooth black curves (*bottom*). Superposed on each process is a rapidly fluctuating force caused by random molecular agitation. The fluctuating force on the grain causes it to undergo Brownian motion about its linear path through the liquid. On the rod the fluctuating force takes the form of temperature fluctuations in small regions of the rod; relatively hot regions are shown in color and cooler regions are shown in lighter color and progressively darker shades of gray. The fluctuations are also graphed as jagged colored lines that oscillate about the smooth black curves; they are exaggerated throughout the diagrams for clarity.

of the number of microstates; the constant of proportionality, which is called Boltzmann's constant, is the universal gas constant R per atom, or 3.3×10^{-24} calorie per degree C.

The upshot is that in statistical thermodynamics the second law can no longer be taken as absolute truth. Since all the microstates of a system are equally probable, there is a small but finite probability that fluctuations will bring about a highly ordered macrostate. More precisely, the probability that a spontaneous fluctuation will lead to a decrease in entropy is proportional to e, the base of the natural logarithms, raised to a power equal to the negative of the change in entropy divided by Boltzmann's constant. For example, the probability of a spontaneous decrease in the entropy of one mole of helium at zero degrees C. is about $1 / 10^{19}$. It is possible, although extraordinarily unlikely, that all the air molecules in a room will spontaneously collapse into a corner, leaving a vacuum in the rest of the room.

The validity of phenomenological thermodynamics is thus a reflection of the fact that by ordinary standards Boltzmann's constant is an exceedingly small number. Could one survive in a world in which Boltzmann's constant was much larger than its present value? Probably not: in such a world the increase in the kinetic energy of each atom for a given increase in temperature would be far greater than it is in our universe. The probability that fluctuations would lead to a reduction in entropy would increase, and the spontaneous appearance of ordered physical systems would be far commoner on a macroscopic scale.

Such a world would be similar to the world one might experience if one could shrink to the dimensions of a Brownian particle. On that scale the pressure, temperature and volume of nearby matter fluctuate constantly. Moreover, near the critical point of a phase transition such as the one from a gas to a liquid the random fluctuations become increasingly large. To express the probabilistic interpretation of the laws of phenomenological thermodynamics the late Maria Goeppert-Mayer and Joseph Edward Mayer borrowed from Gilbert and Sullivan: "What never? No never! What never? Well, hardly ever."

Nonequilibrium Systems

The recognition that Brownian motion is a manifestation of the statistical fluctuations among the microstates of a thermodynamic system had even greater consequences for the study of nonequilibrium systems than it had for systems in equilibrium. At equilibrium the order in which events take place is immaterial: the fluctuations among microstates hardly ever give rise to a distinguishable macrostate. For systems that are not in equilibrium, however, the temporal order of events becomes important.

I have already implied that a statistical version of the phenomenological second law of thermodynamics must explain how nonequilibrium systems evolve to a state of equilibrium. In other words, the second law must explain in detail how ordered macrostates spontaneously give rise to disordered ones. Processes such as the scrambling of an egg are temporally irreversible and so they determine a direction in time. In phenomenological thermodynamics a set of laws based on empirical studies was developed that described the evolution of various irreversible processes from certain special kinds of nonequilibrium state.

For example, if a metal bar is heated at one end and the temperature difference between the two ends of the bar is not too great, the rate at which the heat energy tends to flow from the hot end to the cold end is directly proportional to the temperature difference. Thus as the bar regains thermal equilibrium the "force" driven by the temperature difference decreases and the "flow," or rate of heat transfer, slows down. Similarly, as I have already mentioned, Fick's law holds that the flow rate of a diffusing gas from regions of high to low concentration is directly proportional to the force derived from the difference in gas concentrations between the two regions. Both laws are linear relations because although both force and flow are constantly changing, the flow is always proportional to the force.

In principle nonequilibrium thermodynamics is not restricted to linear relations between forces and flows. The more a system departs from equilibrium, the more important are the nonlinear effects. There is no general way, beyond trial and error, to formulate laws that describe such effects. Nevertheless, a statistical understanding of nonequilibrium systems that obey linear, phenomenological laws might also lead to a better understanding of nonlinear effects. The mathematical treatment of nonequilibrium thermodynamics can be traced to an equation describing the motion of a particle in a viscous fluid that was formulated by the French physicist Paul Langevin in 1908.

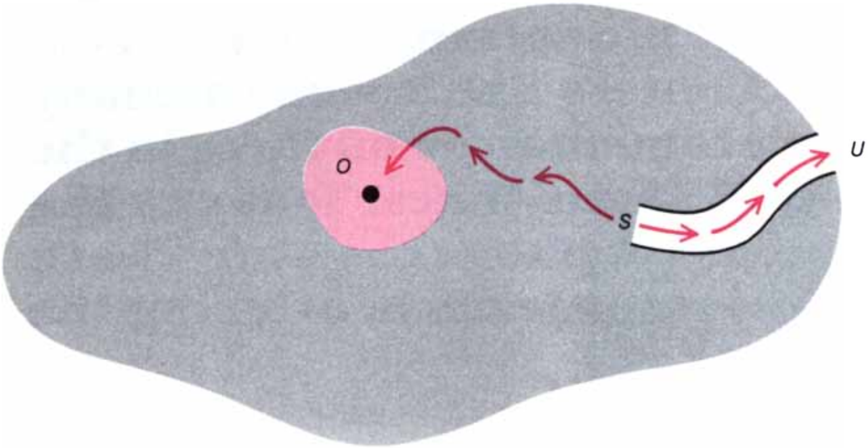
Imagine that a small spherical particle, say the size of a marble, is pulled through a fluid by some outside force. Remember that the particle eventually reaches a terminal velocity that depends on the viscosity of the fluid. The viscosity generates a drag force, proportional to the velocity of the particle, that acts in a direction opposite to the direction of the motion of the particle and tends to slow it down. The energy of the forward motion is dissipated as heat, which gives rise to thermal fluctuations in the fluid.

Imagine now that the size of the particle is a variable and that the size is continuously reduced until it becomes microscopic. A particle of macroscopic size would "feel" only the drag force, and its motion would be essentially unaffected by molecular bombardment. As the particle gets smaller the fluctuations become increasingly noticeable until at last the particle acts as if it were not affected by the macroscopic drag force and undergoes Brownian motion. Langevin's equation combines both effects and applies to a particle of intermediate size.

Langevin's Equation

Langevin's equation for motion therefore has its roots in two worlds: the macroscopic world represented by the drag force and the microscopic world represented by the fluctuating, or Brownian, force. According to the equation, the total force acting on the particle is equal to the sum of the two forces. The two forces can be interpreted as components of the total force that act over different time scales. Over a short time scale the dominant force is the fluctuating force, which varies rapidly. Over a longer time the effects of drag begin to predominate. The total force is equal to the mass of the particle multiplied by the acceleration caused by the two components, which is the rate of change of the velocity of the particle with time.

The Norwegian physicist Lars Onsager made the clever observation that by a mere change in notation Langevin's equation could give a statistical description of an irreversible process. He suggested that the velocity of the particle described in Langevin's equation be replaced by the deviation of a thermodynamic quantity from its equilibrium value. For example, if a rod is heated at one end, its deviation from thermal equilibrium is likened to the velocity of a particle. Furthermore, Onsager proposed that the drag force on the particle be replaced by the drift of a thermodynamic system toward its equilibrium state. Thus the thermodynamic "force" driving the heated rod back to equilibrium is compared with the viscous drag on the moving particle. The resulting equation can then be used to study the influence of thermal



SYMMETRY BETWEEN PAST AND FUTURE is illustrated schematically by the temporal path of a physical system through a series of physical states. Every point within the boundary of the gray region represents a state of the system, and the equilibrium state is at the point *O*. If the system is displaced to some nonequilibrium state *U* that corresponds to the maximum entropy of the system on the boundary, it will almost surely evolve according to the macroscopic laws of thermodynamics. These laws predict that it will pass through some nonequilibrium state *S*, for which the entropy is not less than the entropy at *U*, and finally reach the inside of a small neighborhood of *O* (*color*). Once inside the neighborhood the system will fluctuate almost endlessly about *O*. No matter how small the fluctuations, however, their eventual, cumulative effect is to drive the system back to the nonequilibrium state *S*. The system then makes its way to the point *U* on the boundary. If the system is left to itself, the last leg of the journey from *S* to *U* almost always follows a path within an infinitesimally small cylindrical tube that encloses the path taken earlier from *U* to *S*.

fluctuations on irreversible processes.

Onsager's mathematical trick brings out a deep analogy between the motion of the particle and the decay of a nonequilibrium state. Over a time longer than the time needed for a fluctuation to die out Onsager assumed the average course of decay was given by the phenomenological laws of nonequilibrium thermodynamics. The appearance of a smooth flow of heat from the hot end to the cold end of the heated rod arises in much the same way as the appearance of a smooth drift of a relatively large particle in a liquid. In each case the observations are made over a small but finite interval, which is nonetheless long enough for the random fluctuations to cancel one another. Only over a short time can one detect the fluctuations superposed on the smooth, average drift that is generally observed.

Time-reversed Evolution

Onsager's assumption about the average course of a system toward equilibrium appears to be entirely reasonable, for otherwise how could the linear relations of nonequilibrium thermodynamics emerge from a more detailed statistical description? Nevertheless, it turns out to be restrictive, for it limits his theory to the study of systems whose fluctuations are distributed according to a bell-shaped curve. In other words, the small fluctuations in the temperature of the rod away from the average values given by the phenomenological law of heat transfer must be distributed in the same way as are the displacements of a large number of Brownian particles from a given point at a given time.

What happens if the distribution of the fluctuations is not assumed to be bell-shaped? One can then proceed from the reasonable although less restrictive assumption that any nonequilibrium state of a physical system must drift toward an equilibrium state that is characterized by the vanishing of the drift. According to this condition, the equilibrium state need not be stable in the usual, static sense. Instead the equilibrium state is nothing more than the state toward which all other states tend to be restored. If the system is perturbed from equilibrium, it responds by diminishing the size of the perturbation in such a way that the initial state of equilibrium is restored.

Such a definition of equilibrium recognizes that thermal fluctuations, no matter how small they may be, can alter the evolution of a thermodynamic system. There must always be some finite probability that the system will go against the flow. No matter how far some nonequilibrium state is from the equilibrium state and no matter how small the random fluctuations may be, there is some probability that sooner or later the system will reach the nonequilibrium state.

A detailed mathematical analysis gives an even more remarkable result. It turns out that the approach of the system to the highly improbable, nonequilibrium state tends to follow the reverse of the path the system would follow if it were somehow "placed" in the nonequilibrium state and allowed to drift back to equilibrium, provided there is no interference from outside the system. This antithermodynamic, or time-reversed, evolution is not slow and continuous but takes place instead in leaps and bounds. Given virtually unlimited time, the growth and decay of thermal fluctuations manifests a symmetry in past and future that is oddly contrary to the temporal asymmetry more commonly associated in thermodynamics with the second law.

Mathematical Analogies

The mathematical apparatus developed to treat the fluctuations reflected by Brownian motion can be applied in essentially any discipline in which one would like to estimate the effects of a random variable. Random variables arise frequently in the description of many natural phenomena, primarily because the values of such variables are not known or hard to determine. An early application was the filtering of static, or random electromagnetic noise, from a radar signal or a radio broadcast. By analogy, similar mathematical techniques are useful whenever some kind of "noise" can be identified in a system whose evolution would be determinate in the absence of the noise. The problem is then to estimate the influence of the random variable on the final outcome.

For example, in the study of the dynamics of star clusters the motion of a given star can often be sorted into two components. One component is the gravitational influence of the cluster as a whole, and the other is the gravitational influence of the local stellar neighborhood. The influence of the cluster as a whole varies smoothly over time with the position of the given star; it can be approximated by a simple expression for gravitational potential energy because the number of stars that combine to make up the force is quite large.

The influence of matter in the local stellar neighborhood, however, subjects the gravitational potential of the given star to rapidly varying fluctuations. Such fluctuations make the instantaneous force acting on the star deviate from the force that would result from the smooth gravitational potential energy alone. Analysis carried out in the early 1940's by Subrahmanyan Chandrasekhar of the University of Chicago makes it possible to calculate the probability that the magnitude of the gravitational force on the given star lies between two given values.

In ecological systems there are many cases in which randomly fluctuating variables can influence long-term evolution. Forecasts of population ratios, for example, are based on sample counts of various competing populations that have been made over many years. Such counts vary with the methods employed by the census takers, and the exact points of variation are often difficult to assess. A more reliable forecast can be done by treating such variations mathematically as a fluctuating random variable. Similarly, unpredictable fluctuations of gene frequencies within an animal population can lead to the fixation of a single genotype, and the process can be best understood by treating the fluctuations as random events.

Economics and Navigation

Much more recently probabilistic analyses based on the mathematics of Brownian motion have been applied in economics. Economists have applied such techniques to describe the behavior of stock prices, inflation rates, spot rates of interest and other financial variables. For example, the price of an asset is controlled in part by the trading of options, which are contracts giving the right to buy or sell the asset within a specified period. The price of the asset can be expected to fluctuate according to the number and the price of options being traded, and the fluctuation is superimposed on the price that would presumably be determined by pure market forces in the absence of options. The aim, as always, is to predict the future price of the asset with the highest confidence possible.

Given an estimate of the effects of a random fluctuation, one would like to control it. Again the mathematical techniques developed in one discipline have been found to generalize quite readily. Statistical control similar to that used for improving the reception of radar signals during World War II is now employed by many systems that sample continuous streams of data, such as radio-navigation systems. It has also been applied in the control of the quality of mass-produced goods.

Although the physical Brownian motion that gave rise to the mathematical theory is now well understood, the development of mathematical analogies is by no means complete. The history of such applications suggests that Brownian motion will continue reappearing, in progressively abstract form, to illuminate various disciplines for many years to come.

Reproductive Success in Red Deer

A 12-year study of more than 1,000 of these gregarious animals in a nature reserve of the U.K. has revealed the determinants of the lifetime breeding success among red deer stags and hinds

by T. H. Clutton-Brock

In a remote glen on the Scottish island of Rhum a red deer stag stands among his hinds. He is in his 10th year, and the large harem that surrounds him on the close-cropped turf is likely to be the last he will hold. It includes hinds of several ages. Two matriarchs graze at opposite sides of the group. They are aged 15 and 13. Each of these old hinds presides over a large family of mature daughters with whom they share the two square kilometers of moorland, hillside and valley bottom that make up their home range. Between the matriarchs stand their eight daughters, three with sleek coats and rounded hindquarters and no calf at foot. Five are with calves, the angular silhouettes of the mothers bearing witness to the heavy cost of lactation. Two three-year-olds, which will become pregnant for the first time this year, graze the sides of a slight depression in the dunes. Eight two-year-olds, yearlings and calves are dispersed throughout the group. Some 200 yards away three yearling males, recently separated from their maternal groups by the harem master, graze uneasily, not yet altogether accustomed to independence.

On the knolls above the harem a six-year-old stag appears, black from rolling in the peat bogs. He paces purposefully toward the harem and then stops and gives a series of coughing roars more like a lion's roar than the bugling of the American elk. The harem master replies, and for some 15 minutes the two stags roar at each other repeatedly. Then, with slow and deliberate steps, the challenger moves closer to the harem and the master stag moves out to meet him. They almost touch, but then, as if by mutual consent, they turn into a parallel walk, pacing tensely a dozen meters apart.

Back and forth they go until the resident stag, identifying a potential advantage, runs up on a small hillock, lowers his head and charges down on the challenger. The six-year-old immediately lowers his head to bring his antlers into play, and the two stags meet with the dry click of antler on antler. Both brace their front legs forward, shoving hard with straining haunches. They circle once. The young stag suddenly leaps forward, trying to thrust the bigger stag sideways, but the resident braces his back legs and slowly forces his opponent's head up. With a quick twist he turns the younger stag's head, throws him sideways and charges at his exposed flank. The challenger turns and flees. For a few yards the winner pursues; then he stops and roars at his retreating rival before returning to his hinds. They have shown no interest in the struggle and have continued to graze methodically. The master will keep the harem at least until the next challenge, but at his age his fighting ability is already waning. Next year the contest would probably have a different outcome.

This interaction emphasizes a fundamental difference between males and females that is found in virtually all mammals where successful males monopolize breeding access to several females. Because approximately equal numbers of males and females are born, the ability of some males to hold harems of from 10 to 20 females necessarily means that within any given year a large proportion of males will not be able to breed. Hence competition for harems is intense and males fight regularly for possession.

The situation is different for females. A hind's breeding success is not directly related to the number of mating partners. Instead it depends on the number of offspring the hind can produce and rear in her lifetime. Because the costs of gestation and lactation are heavy, a female's success is likely to depend mainly on the food resources available to her.

The difference between males and females in the nature of competition when males defend harems indicates that the characteristics affecting breeding success must also differ between the sexes. Fighting ability, large body size and highly developed weapons will be important to males. In contrast, there may be little advantage to females in fighting for widely dispersed food supplies. Instead their success is likely to depend mainly on the efficiency with which they extract food from their environment and convert it into offspring. This conclusion need not imply that their ability to displace other females from feeding sites is unimportant, but it does imply that the ability to win fights is unlikely to have as strong an effect on their breeding success as it does on the success of males.

The differences in the determinants of breeding success in males and females are likely to be sharper in polygynous species (where males monopolize breeding access to several females) or in polyandrous ones (where females monopolize breeding access to several males) than in monogamous species. In the monogamous species competition among males and among females is likely to assume more similar patterns.

The contrasting factors affecting reproductive success in males and females are the ultimate cause of most differences between the sexes. Except among animals that breed cooperatively, selection will favor traits that increase the individual's breeding success. Where a trait has a greater effect on breeding success it is likely to be more highly developed in that sex. For example, in species where body size affects the breeding success of males more than that of females and the costs of large size are similar for the two sexes, males will be larger than females. The phenomenon is seen in red deer and elks, where males are almost twice as heavy as females. This argument also predicts that sex differences

RED DEER STAG is accompanied by his harem of hinds on the Scottish island of Rhum in the Inner Hebrides. The breeding success of red deer males increases with the size of their harem. Harem size depends on the stag's fighting ability: a stag that loses a fight with a challenging stag typically loses his harem and may not be able to regain it. One calf is visible in the middle of this group. The five hinds are foraging on seaweed along the shore of the island. The size of a stag's harem can range from one hind to as many as 20.

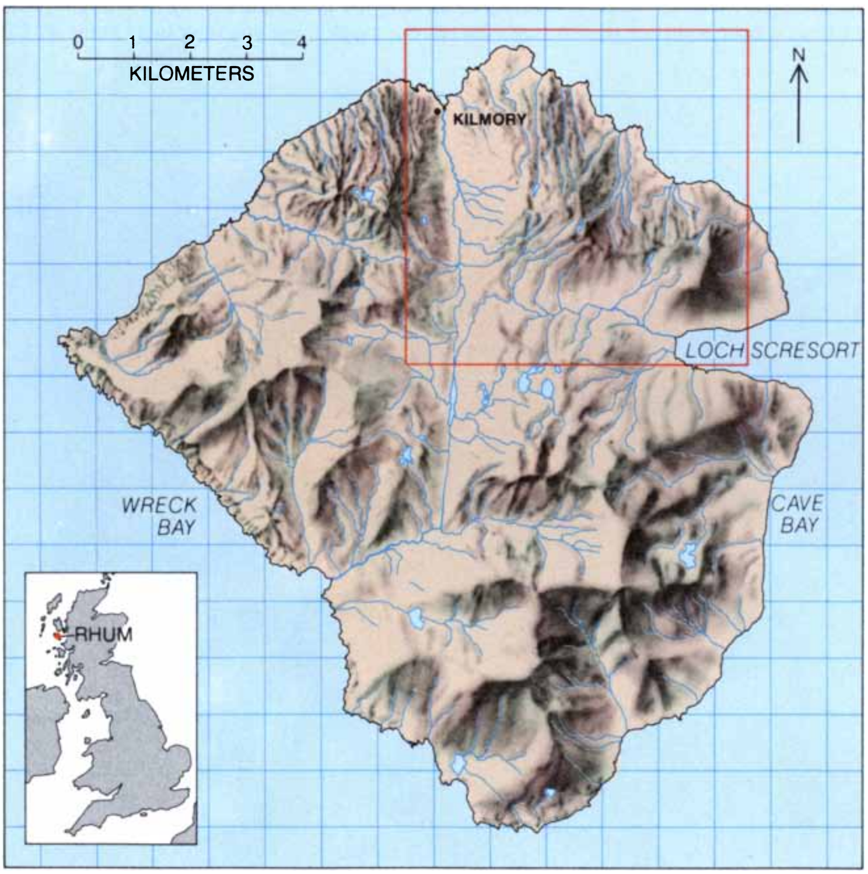
HAREM MASTER responds to a challenge from a younger stag by roaring. If the challenger continues to approach, he and the master may perform a parallel walk so that they can assess each other at close range. Approximately 40 percent of such walks lead to fights.

generally should increase with the degree of polygyny or polyandry.

If differences in the determinants of reproductive success lie at the root of most sex differences in morphology and behavior, an understanding of the factors affecting breeding success in males and females should help to predict and explain differences between the sexes. To work out the causes of variation in breeding success in males and females is not easy, however. What matters, of course, is the individual's success over its entire lifetime. Yet even the number of offspring reared may not be a good measure of success, since offspring born in large litters may have a lower chance of breeding successfully. The number of grandchildren would be a better measure of an individual's success, but the logistic problems associated with ascertaining that number among wild animals are enormous. There are not many species in which it is possible to measure even the breeding success of individuals throughout a particular season, let alone the number of each individual's grandchildren.

An exception is provided by the Scottish red deer (*Cervus elaphus*), which are usually regarded as belonging to the same genus as the American elk. Both sexes are easily visible in the virtually treeless habitat of the Highlands. With practice an observer can learn to recognize individuals by facial structure and coat pattern. The deer have a short and predictable mating season in October, when the stags defend well-defined harems. The calves, born the following June, can usually be caught, marked and weighed. Moreover, both sexes have life spans short enough (about 10 years for stags and 12 for hinds) for human observers to be able to follow individuals throughout their life.

These are the reasons my colleagues Fiona Guinness, Stephen Albon and I at the University of Cambridge have spent the past 12 years following individuals on the Inner Hebridean island of Rhum. At any one moment our study area has about 300 red deer, but over the course of the study more than 1,000 have passed through our population. In recent years the calves we caught and marked in the early part of the study have begun to complete their life spans, and a picture is now emerging of what determines breeding success in males and females.



ISLAND OF RHUM is the site of the 12-year study of red deer that has been made by the author and his colleagues. The island, off the northwest coast of Scotland, is a National Nature Reserve of the U.K. Study area is the northeastern section of the island (*red box*).

That success varies more among the stags than among the hinds. The number of calves a stag fathers per year ranges from none to about 10. (It falls to six when only the calves that survive for at least a year are considered.)

The differences in success would be quite large if these contrasts were maintained throughout a stag's life, but they are not. A stag's breeding success is strongly affected by his age and usually peaks between the ages of seven and nine years. At any given time one can estimate a stag's success by the size of his harem, but the variation in lifetime breeding success is smaller than the large differences in harem size within a season would appear to suggest. Even the most successful stags seldom father more than two dozen surviving offspring during their life.

In contrast, the variation in breeding success among hinds is greater than one might suppose from the fact that a hind has at the most one calf per year. The reason is that the same hinds tend to lose their calves throughout their breeding period (from six to 12 years). The number of surviving calves raised by females during their lifetime ranges from none to 10 and is closely related to the number of their grandchildren.

An individual's breeding success is made up of three components: the life span, the number of offspring produced per year and the proportion of offspring that survive. The relative importance of these components differs markedly between the sexes. A hind's success depends partly on living a long time but mostly on the proportion of her calves that survive. Some hinds produce a calf per year throughout most of their life but never manage to rear even one successfully.

Among stags, on the other hand, differences in breeding success depend almost entirely on the number of offspring they sire per year between the ages of six and 11. The differences in survival among calves sired by different stags are small, and the duration of a stag's life beyond the age of 11 has little effect on his lifetime breeding success.

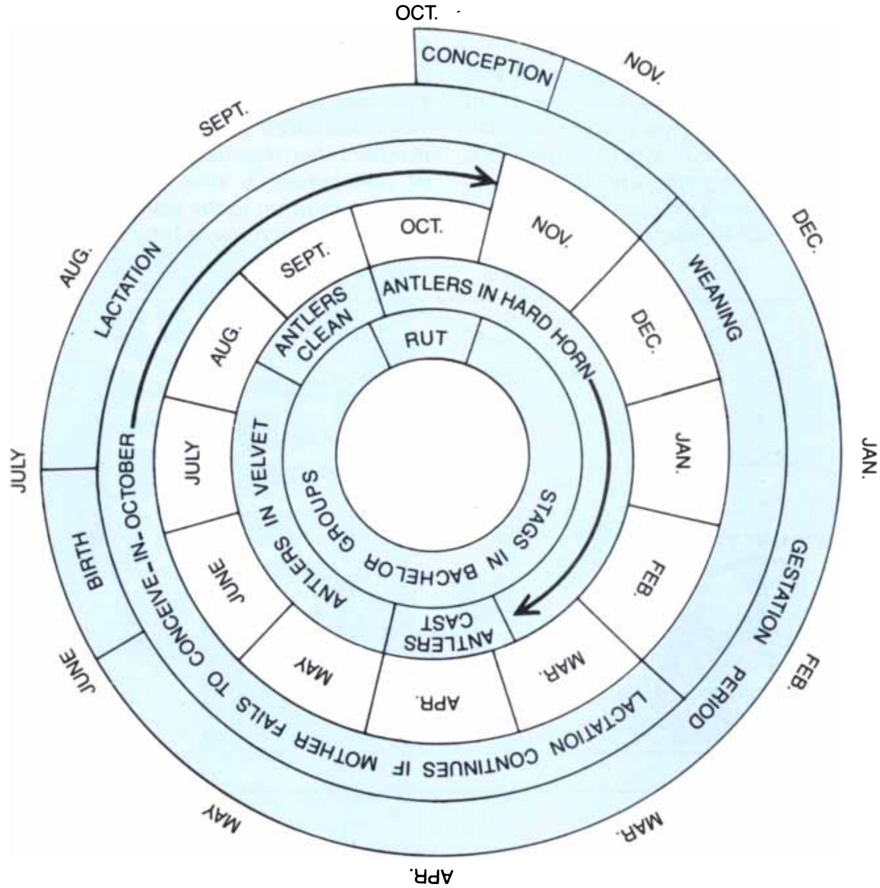
The factors that affect breeding success also differ between stags and hinds. The ability of a female to rear her calves is influenced mainly by the quality of her home range (which affects her access to food resources), the number of animals with which she shares it (which determines the intensity of competition for food) and her social dominance (which affects her priority of access to preferred feeding sites). A hind inherits a home range from her mother and shares it with her sisters and other matrilineal relatives, so that much of her breeding success depends on the identity and size of her family. Her dominance rank is related to her birth weight and her adult weight: heavy calves grow rapidly and develop into large, dominant females.

In contrast, the breeding success of stags depends almost entirely on their fighting ability. That in turn is related to their body size and to their growth during the first year of life. In red deer, as in many other species where males are substantially larger than females, male growth is more severely affected by a shortage of food than female growth. Partly for this reason and partly because success in breeding varies more widely among males than among females, the breeding success of stags is even more strongly influenced by the dominance and body size of their mother than the success of females. As a result the sons of dominant hinds are more successful than the daughters, whereas the daughters of subordinate hinds outperform the sons.

The contrasting factors affecting the breeding success of red deer stags and hinds provide important insights into the evolutionary origins of many sex differences in morphology, physiology and behavior. The reasons for some sex differences are obvious. The fact that size and fighting ability are so important to red deer stags indicates why they are nearly twice as heavy as hinds and have developed such elaborate weaponry in the form of antlers. As this explanation predicts, sex differences in body size are greatest in deer species where successful males hold large numbers of females, as reindeer and elks do. They are least in species such as roe deer and muntjacs, where the opportunity for males to breed polygynously is small because adult females are widely distributed.

The contrasting factors affecting the breeding success of males and females also shed light on the origin of less obvious differences between the sexes. In many mammal species (including red deer) juvenile males grow faster than females but lay down less body fat, forgoing the advantages provided by reserves of fat for the advantages of rapid early growth and large adult size.

These differences too vary systematically across species. Sex differences in juvenile fatness are most pronounced in polygynous species (in which other physical differences are also greatest). In less dimorphic species juvenile fatness is about the same in males and females. Among Chinese water deer—the one species of deer whose females are larger than males—the males are fatter as juveniles.



REPRODUCTIVE CYCLE of the red deer on Rhum is charted. The outer two rings depict the cycle for hinds. The inner two rings relate to the cycle for stags. The breeding season is in the fall, and more than 80 percent of the hinds conceive in October. By beginning to follow the hind's cycle in October (*top*), one sees that the calf is born in June, as most red deer calves are. Most of the calves are weaned six or seven months later if the mother conceives again but not until about 18 months after birth if the mother fails to conceive in the second fall. The outer ring of the cycle for stags shows the monthly state of the antlers. The inner ring shows the clearly defined breeding season of the red deer, which is mainly in October.

Such male phenomena as larger size, faster growth rates and reduced fat reserves must have their costs, otherwise differences between the sexes would increase continually. In red deer and many other sexually dimorphic mammals males have shorter lives than females and are more likely to die of starvation during periods of food shortage or harsh weather. One of the most spectacular examples on record was in the reindeer population of St. Matthew Island in the Bering Sea: a population of more than 6,000 adults crashed to 42 during the course of a single winter, and only one of the survivors was a male.

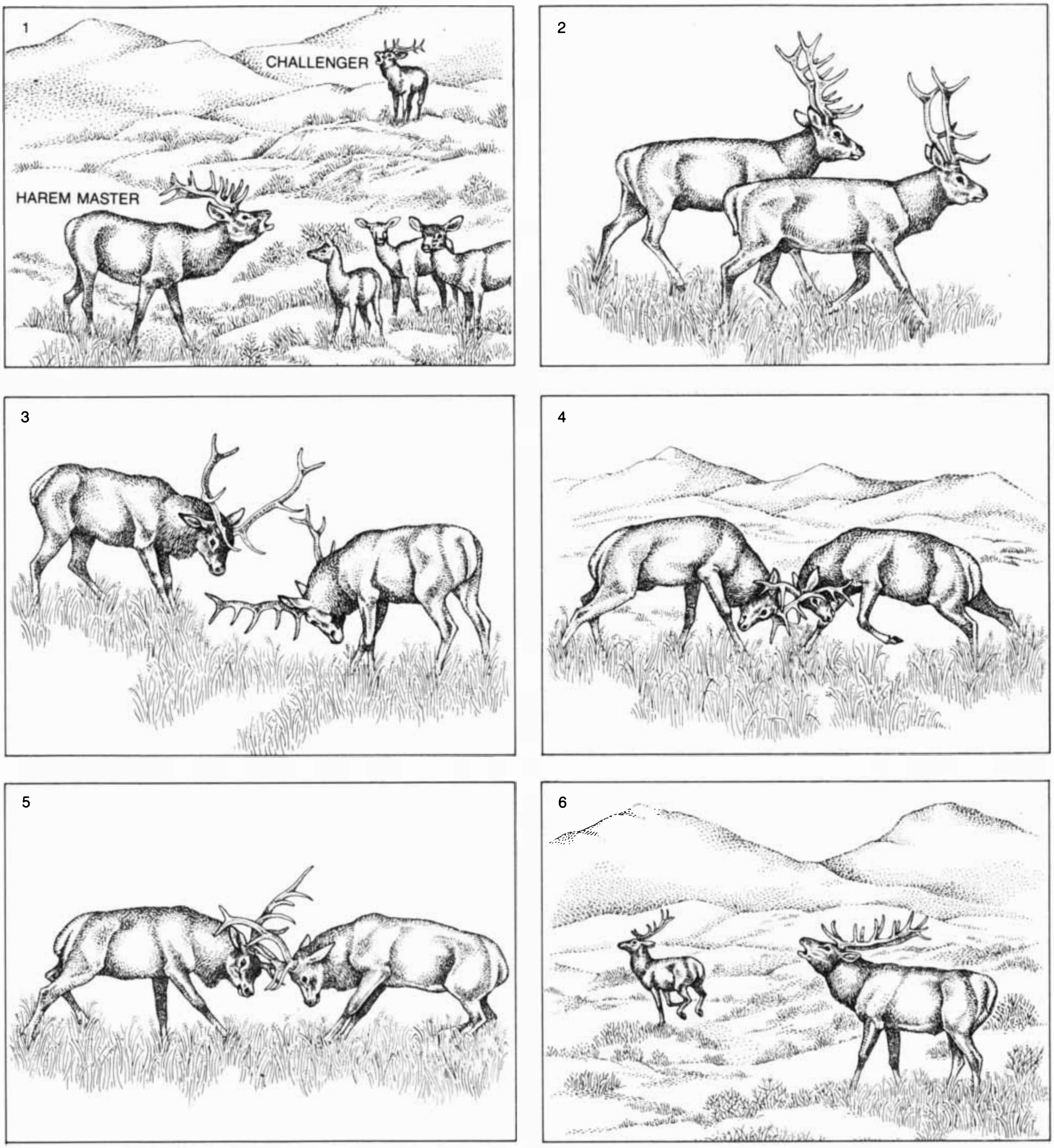
Sex differences in survival are also found among juveniles. Male red deer as calves and yearlings are much more likely than females to die during the winter. In the early years of our study, when the population density was fairly low, no sex differences in juvenile mortality were apparent, but as density rose the situation changed. Now some 60 percent of the males die during their first or second winter. For females the rate is 30 percent.

Sex differences in the effects of early growth on breeding success suggest that mammalian mothers might be expected to treat sons and daughters differently. The fact that early growth and maternal investment in the form of milk are likely to influence a male's breeding success suggests that mothers would increase the number of their grandchildren by devoting a greater proportion of their reproductive effort to their sons than to their daughters. There is some evidence that they do so. Male red deer calves are born about half a kilogram heavier than females. Similar differences are found in species that produce litters.

Inasmuch as growth during the first weeks of life is likely to affect the breeding success of males more than that of females, mothers might be expected to give more milk to their sons. Male red deer calves on Rhum suckle longer and oftener than females. Higher suckling rates by juvenile males have been found in several other polygynous mammals.

Our comparison of hinds according to whether they had raised sons or daughters shows that rearing sons entailed the greatest costs in terms of the mother's subsequent reproductive success. In any season about 25 percent of the hinds fail to bear calves. Failure to breed is largely a consequence of poor body condition in the fall. In our population 33 percent of the mothers that rear sons fail to breed in the following year, compared with 18 percent of the mothers that rear daughters. The date of conception is also related to the hind's condition in the fall. Hinds conceive nearly two weeks later in the year after raising a male than in the year after raising a female.

The argument that mothers should invest more heavily in sons to maximize the number of their grandchildren does not necessarily apply to all the individuals in a population. As Robert L. Trivers of the University of California at Santa Cruz originally



CHALLENGE TO HAREM MASTER begins when a young stag approaches him (*1*). The two stags usually roar at each other for several minutes. If the challenger moves closer, the master goes out to meet him. In most cases the next move is a parallel walk (*2*). Both stags pace back and forth together, a few meters apart, for several minutes. If the challenger does not withdraw, one of the stags initiates a fight by turning and lowering his antlers (*3*). They lock antlers and push (*4*). In this encounter the master finally wins by forcing the challenger backward (*5*). In the end the challenger flees and the master pursues for a few yards, continuing to roar at him (*6*).

pointed out, inferior mothers whose poor physique makes them unlikely to raise successful sons might be expected to desert their male progeny soon after birth, conserving their resources for subsequent attempts at breeding. For example, subordinate hinds whose sons are unlikely to breed successfully might withhold milk from them.

Is this the case? If it is, the sons of subordinate mothers should be more likely than the daughters to die, whereas no sex differences in survival should be found between the sons and daughters of dominant animals. The red deer data show precisely this pattern: 51 percent of the sons born to subordinate mothers die within their first two years of life, compared with 30 percent of the daughters. No similar sex difference is found between the sons and daughters of dominant mothers: 38 percent of their sons die within their first two years of life, compared with 36 percent of their daughters.

Yet to desert offspring after they are born carries a heavy cost, since each one represents a large fraction of a female's lifetime breeding potential. Mothers could increase their breeding success by manipulating the sex ratio before birth, ideally as close to conception as possible. The relation between female rank and the breeding success of male and female offspring leaves little doubt that dominant mothers would increase the number of their grandchildren by giving birth to more sons than daughters and that the reverse would be true for subordinates.

Does this happen? We compared hinds allocated to three equal categories of rank. The comparison showed that the sex ratio of calves lies at about 47 percent male for subordinates, 54 percent for middle-ranking hinds and 61 percent for dominant hinds. The differences are consistent in all eight cohorts of hinds for which we have breeding data covering at least 75 percent of the life span. Furthermore, across individual mothers the sex ratio increases with the mother's rank from about 30 percent male calves to about 70 percent for the most dominant hinds. The rank at which individuals exceed the average sex ratio for the population coincides closely with the point at which sons are more successful than daughters. A study of rhesus monkeys in Puerto Rico has yielded similar findings.

These are surprising results, because research on a wide variety of domestic birds and mammals has found that it is impossible to manipulate the birth sex ratio of progeny by selective breeding. One possible explanation is that a consistent sex-ratio variation can arise as a

YEARS	HINDS BORN IN 1972		
	FIRST INDIVIDUAL (THII)	SECOND INDIVIDUAL (MOMC)	THIRD INDIVIDUAL (COLC)
1972	THII (CALF)	MOMC (CALF)	COLC (CALF)
1973	THII (YEARLING)	MOMC (YEARLING)	COLC (YEARLING)
1974	THII (TWO YEARS OLD)	MOMC (TWO YEARS OLD)	COLC (TWO YEARS OLD)
1975	NO OFFSPRING	NO OFFSPRING	CALF SURVIVES
1976	CALF SURVIVES	CALF (DIES IN SUMMER)	CALF (DIES IN SUMMER)
1977	CALF SURVIVES	CALF (DIES IN SUMMER)	CALF SURVIVES
1978	THII DIES	CALF (DIES IN SUMMER)	CALF SURVIVES
1979		CALF (DIES IN SECOND WINTER)	CALF SURVIVES
1980		CALF (DIES IN FIRST WINTER)	CALF SURVIVES
1981		NO CALF	CALF (DIES IN SUMMER)
1982		CALF (DIES IN SECOND WINTER)	CALF SURVIVES
1983		CALF (DIES IN FIRST WINTER)	CALF SURVIVES
1984		NO CALF	NO CALF
TOTAL SURVIVING OFFSPRING	2	0	7

BREEDING SUCCESS OF HINDS is charted for three of the many females born on the island of Rhum in 1972. A comparison of Momc with Colc shows how breeding success, measured in terms of calves that survive to adulthood, can vary between red deer hinds. Social rank is a factor in breeding success: Colc was relatively dominant, Momc subordinate.

response to individual differences in the size or condition of the mother's body in populations where food is limited. Perhaps consistent trends are not found in domestic animals because they are maintained on an artificially high plane of nutrition, so that differences in condition are small. Alternatively, the ability to vary the sex ratio in response to environmental changes may have been lost in domestic animals as a result of artificial selection over many generations.

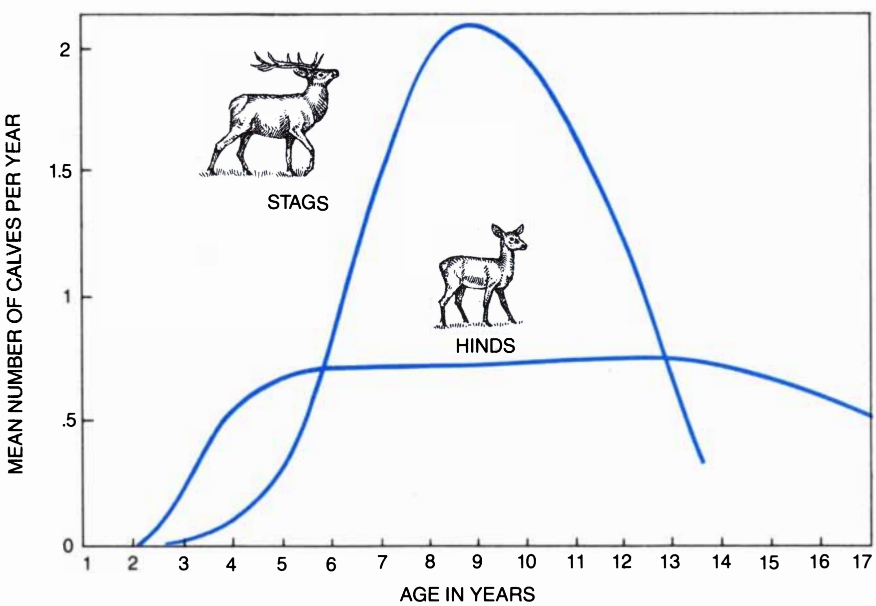
Our research does not yet show whether sex-ratio variation resulting either from sex differences in mortality after birth or from variation in the birth sex ratio is controlled by the mother. One possibility is that young males are generally more susceptible than young females to food shortage. If they are, it would explain why the sons of subordinate mothers are more likely to die than the daughters. Comparisons of suckling between the sons and daughters of subordinate hinds have so far yielded equivocal results.

Another possibility is that subordinate hinds may give birth to fewer sons than dominant hinds because their male offspring are more likely to die before birth. Studies of several different mammals show that the sex ratio of aborted fetuses has a male bias.

A reduced survival of male fetuses, however, would explain our results only if all hinds conceived a male-biased sex ratio. As it is, dominants produce significantly more than 50 percent males. Moreover, if the observed

YEARS	STAGS BORN IN 1972			
	FIRST INDIVIDUAL (BROC)		SECOND INDIVIDUAL (YHOC)	
	MEAN HAREM SIZE	CALVES SIRED	MEAN HAREM SIZE	CALVES SIRED
1972	CALF		CALF	
1973	YEARLING		YEARLING	
1974	TWO YEARS OLD		TWO YEARS OLD	
1975	THREE YEARS OLD		THREE YEARS OLD	
1976	NO HAREM	NO CALVES	ONE HIND	NO CALVES
1977	NO HAREM	NO CALVES	FOUR HINDS	FOUR CALVES
1978	ONE HIND	NO CALVES	THREE HINDS	TWO CALVES
1979	NO HAREM	NO CALVES	FOUR HINDS	TWO CALVES
1980	THREE HINDS	ONE CALF	FIVE HINDS	THREE CALVES
1981	TWO HINDS	NO CALVES	FIVE HINDS	TWO CALVES
1982	NO HAREM	NO CALVES	DIED	
1983	DIED			

BREEDING SUCCESS OF TWO STAGS is charted in terms of harem size and the estimated number of calves sired per year. Harem size depends on a stag's ability to win and hold his harem in fights against all challengers. Fighting ability peaks at age seven to 10.



AGE influences the breeding success of red deer stags; the fighting ability of a stag increases as he gets older but then declines when at the age of 10 he passes his prime. Age is less important among red deer hinds, where breeding success does not depend on fighting ability.

trend appears because subordinate mothers lose male fetuses during gestation, subordinates should produce fewer calves than dominants. Differences in fecundity between dominant and subordinate hinds are too small to account for the difference in sex ratios.

Our records of the life histories of red deer on Rhum raise many questions about the causes of reproductive trends in mammals that will be answered only by controlled experiments under laboratory conditions. Why are males more susceptible than females to food shortages? Do mothers really respond differently to male and female offspring, or could differences in the rate of suckling be a result of sex differences in demand? What mechanisms are responsible for the consistent male bias in birth sex ratios found in many mammals? Do small, subordinate hinds conceive more sons than dominant hinds?

We are not likely to produce final answers to any of these questions on Rhum. Nevertheless, our research illustrates the role that field studies of the reproductive strategies of animals in their natural environments can play in suggesting the direction of research in related fields. To cite one example, it seems probable that attempts to manipulate the birth sex ratio experimentally might have been more successful if they had been carried out on animals showing a pronounced sexual dimorphism and maintained in conditions where food was limited.

In order to test many of the functional explanations of sex differences that have been suggested, future studies will need to examine the determinants of breeding success in species that do not conform to the general trends. For example, zebras and spotted hyenas, like red deer, breed polygynously, but there is little difference in size between male and female zebras, whereas female spotted hyenas are larger than males. Is the effect of body size on breeding success similar in male and female zebras? Does size affect the success of female spotted hyenas more than that of males?

Studies of species with contrasting breeding systems are also needed. Are the determinants of breeding success in both sexes usually more alike in monogamous species than they are in polygynous ones? How do they differ in polyandrous species? The challenge facing field biologists investigating these problems will be to find species in which it is possible to measure the breeding success of both sexes throughout their natural life span and to investigate the causes of difference between individuals.

A FEW GOOD REASONS TO BUY ATHENA, THE NEW PERSONAL COMPUTER FROM CANON.

Athena runs all popular IBM PC compatible software.

If you're in the market for a smart, fast, modestly priced computer that's compatible with the IBM PC, look no further. Canon has just created the ideal one—the new Athena.™

Named for the legendary goddess of knowledge, we believe Canon's Athena is going to create a legend of its own. For one thing, it can run hundreds of programs designed for the IBM PC. No modification is required. (Even the layout of the keyboard is compatible with the IBM PC.)

And with its 8086 16-bit microprocessor, memory board and display adapter, Athena can run the programs even faster than IBM's PC.

Athena can expand as you expand.

Athena offers you five expansion slots. A printer interface, asynchronous communication port and mini floppy disk controller are built in. So only one slot is taken up, with a display adapter. That leaves four expansion slots to use for optional hardware. And with Canon's memory expansion board, you can increase the standard 256K bytes to an enormous 512K bytes.

You also have the choice of either a monochrome or color graphic display unit; each with pedestals for easy adjustment. And, when you're ready, Canon offers a full line of high-quality printers designed to meet any need.

Athena comes with Canon quality as a standard feature.

Perhaps the best feature Athena has to offer can be summed up in a single word—Canon. Our worldwide reputation for quality and dependability has been earned in everything from sophisticated cameras to hard-working office machines.

So you know you're not only getting a great computer, you're getting a great company to stand behind it. How much does all of this cost? A lot less than you'd expect. For more information about Athena, see your local computer dealer or call 1-800-323-1717, x325. (In Illinois, 1-800-942-8881, x325). Dealer inquiries welcome.

Canon

Canon USA, Inc., One Canon Plaza, Lake Success, N.Y. 11042.

Phased-Array Radars

Such a radar can track or search for objects without moving its antenna. To steer the beam it relies on wave interactions among signals from a multitude of small antenna elements

by Eli Brookner

The ceaselessly turning radar dish, sweeping its beam of microwave radiation along the horizon in search of distant objects, is a staple of motion pictures and, in the form of airport radar, of everyday experience. Yet in many of the most familiar uses of radar, such as aviation, air defense and intelligence, the mechanically steered dish is giving way to a new kind of device. A flat bank of small, identical antennas, each one capable of transmitting and receiving signals, takes the place of the concave reflector, and even as its beam scans expanses of sky the radar itself does not move. Instead the signal is deflected from target to target electronically, steered through the principle of wave interference. This new technology is known as the phased array.

The basic concepts of radar remain unchanged in their new embodiment. All radars function by emitting directed beams of radio energy. Most often the energy falls within microwave bands, from 300 million to 10 billion hertz, although some very-long-range radars operate instead in the high-frequency (HF) and very-high-frequency (VHF) bands, respectively from three million to 30 million hertz and from 30 million to 300 million hertz. Depending on the shape of the antenna, the radiation forms a narrow, pencil-like beam, suited for precise tracking, or a dispersed, fanlike beam, best for surveying broad areas of sky.

When the beam strikes an object, it reflects. Provided that the energy of the transmitted pulse, the sensitivity of the antenna and the reflective quality of the object are all sufficient, a detectable echo will return to the antenna. Depending on the type of radar and the kind of pulse transmitted, the echo yields various kinds of information.

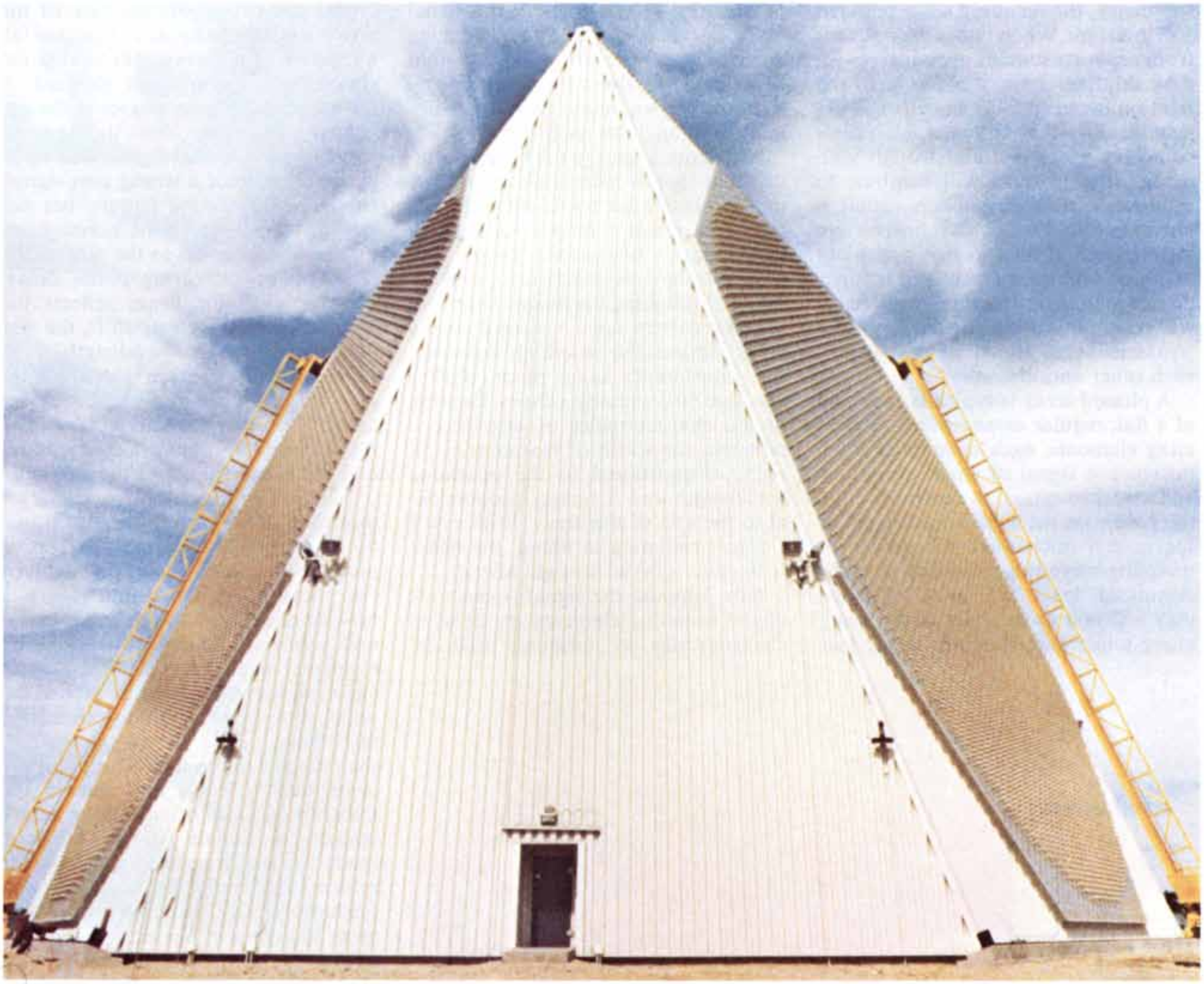
The direction from which the echo returns gives the bearing of the object, and if the radar transmits pulses of energy rather than a continuous signal, the lag between the transmission of a pulse and its echo indicates the object's distance. Some radars are also designed to gauge the Doppler shift of the echo: the change in the frequency of a signal that occurs when the source (in this case the target) and the receiver (the radar installation) are moving with respect to each other. From the Doppler shift such radars derive the object's velocity toward or away from the antenna.

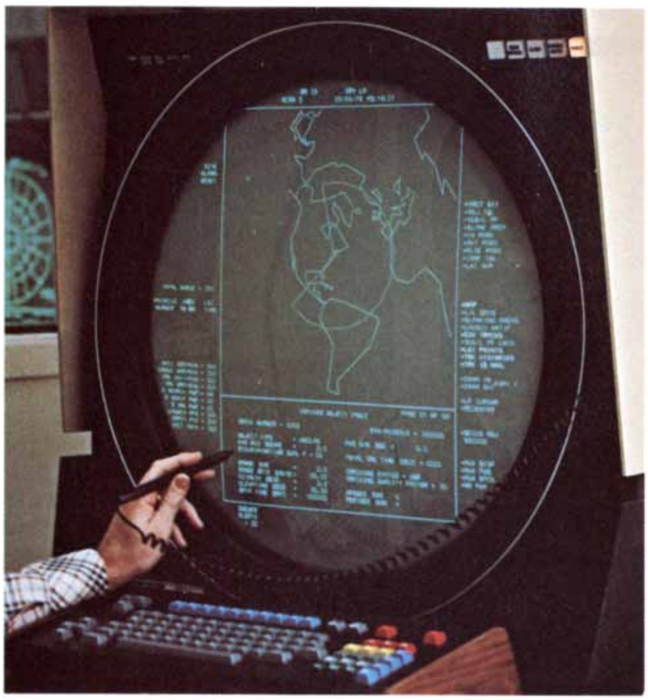
For a given distance the strength of the echo gives some indication of the object's size. The word "indication" is used advisedly; two objects of the same size, if they are shaped differently or made of different materials, will return echoes that differ sharply in strength. To get more precise information on size some radars transmit pulses so brief that they are physically shorter than the targets they are likely to strike, or "illuminate." If a radar emits energy for only a few billionths of a second, the front of the pulse will have traveled no more than a few feet by the time the transmission is complete. Such a pulse is shorter than, say, an airplane. Radar signals reflect from both the near and the far end of their target; in the case of an extremely short pulse the two reflections will result in distinct echoes. The interval between the echoes corresponds to the length of the target.

As a conventional radar surveys broad areas of sky it can gather information about a large number of objects. Inevitably, however, there is a substantial lag between successive observations of individual targets. The update rate, the rate at which a radar takes new readings of a target, is for most mechanically steered radars no faster than the rate at which the radar dish turns on its shaft. On an air-traffic-control radar, for instance, the radial green stripe that sweeps around the cathode-ray-tube display, leaving behind it updated positions and other information about the aircraft within the range of the radar, turns at the same rate as the physically rotating radar dish. The update rate of such radars is typically only about once every six seconds, and even advanced military radars rarely achieve update rates greater than twice a second.

There are circumstances that demand more frequent readings of target position and movement. A single mechanically steered radar can provide continuous data on one or a few closely spaced objects by tracking them, rotating to match their movement. For many military and intelligence purposes, however—shipboard tracking of several missiles attacking from different directions, for example, or the close scrutiny of a number of separate payload components in the test of an intercontinental ballistic missile—each of a large number of targets must be watched continuously. Until recently only groups of radars, each assigned to one or several of the targets, could serve that purpose. Today a single phased-array radar can do what previously might have required a battery of mechanically steered dishes. To give a concrete example, a phased array code-named the COBRA DANE, which observes Soviet ballistic-missile tests from its site on the edge of the Bering Sea, can follow hundreds of targets scattered through a volume of space spanning 120 degrees in azimuth and about 80 degrees in elevation. In effect it watches them simultaneously, electronically redirecting its beam from target to target in a matter of microseconds (millionths of a second).

The electronic beam steering that underlies such remarkable capacities takes advantage of a simple physical principle. When adjacent sources radiate energy simultaneously at the same





PAIR OF PHASED ARRAYS make up a Janus-faced radar at Otis Air Force Base on Cape Cod (*top*). Known as the PAVE PAWS and built by the Raytheon Company, the installation is designed to provide early warning of attacks by submarine-launched ballistic missiles and to aid in tracking satellites. In contrast to the steerable dishes of conventional radar, no part of the PAVE PAWS antenna moves. Instead its beams are steered electronically by the interference of the separate signals emitted by 1,792 radiating elements on each 102-foot-wide face. The twin antennas, facing 120 degrees apart, each capable of scanning 120 degrees of azimuth, combine to give the radar the 240-degree field of view outlined on its computer-synthesized display (*left*). Within that angle it can follow a large number of objects almost simultaneously by electronically shifting its beam from one target to another within a few millionths of a second. At its 3,000-nautical-mile range the PAVE PAWS installation can detect objects with a frontal area of 10 square meters. The light pen enables an operator to summon data about the targets to the screen and to change the information in the data files.

frequency, the resulting wave patterns will interfere. When two wave patterns from separate sources are superposed, their relative phase, a measure of the relation between their waveforms, determines how they will interact. If crest coincides with crest and trough with trough, the patterns will combine to produce a strengthened sum signal, a phenomenon known as constructive interference. If the waveforms are out of phase and their crests and troughs do not match, destructive interference will result: the two signals will yield a weakened sum signal or will cancel each other entirely.

A phased array is typically made up of a flat, regular arrangement of radiating elements; each element is fed a microwave signal of equal amplitude and matched phase. A central oscillator generates the signal; transistors or specialized microwave tubes such as traveling-wave tubes amplify it. If the signals all leave the array in phase, they will add up in phase at any point along a line perpendicular to the face of the array. Consequently the signal will be strong, capable of producing detectable echoes from objects that lie in its path, along the array's boresight, or perpendicular axis, and within a small angle to each side.

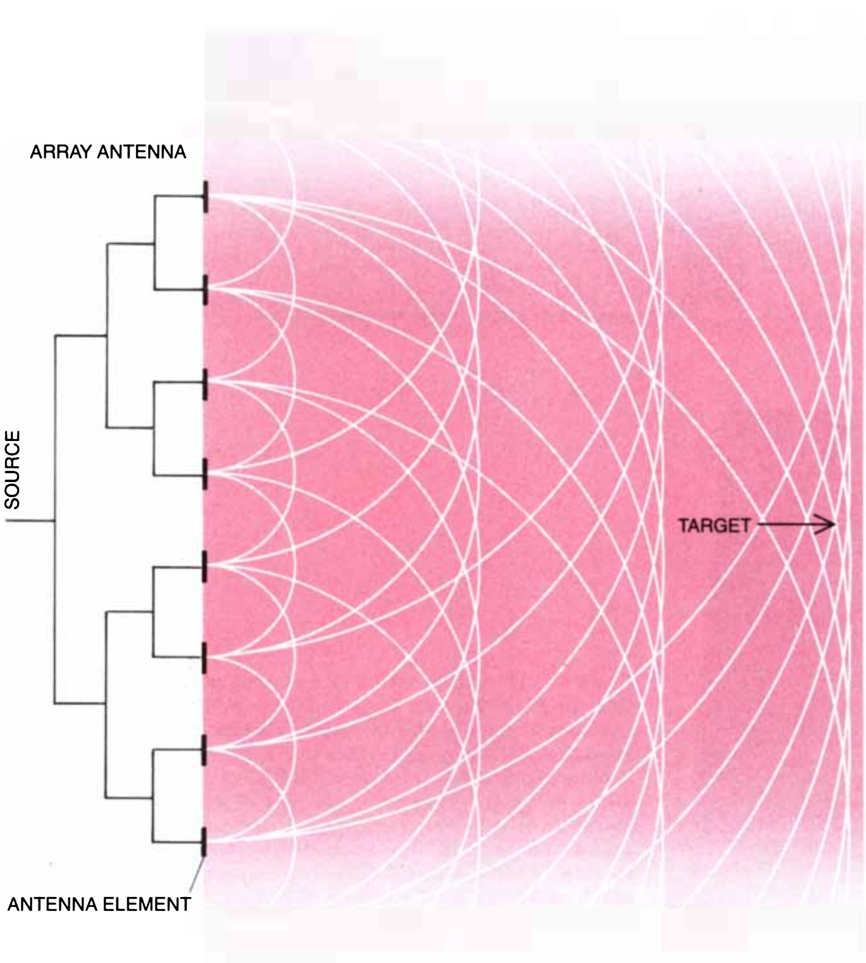
At greater angles to the boresight individual signals from different radiating elements must travel different distances to reach a target. As a result their relative phases are altered and they interfere destructively, weakening or eliminating the beam. Thus outside the narrow cone, centered on the array's boresight, in which constructive interference takes place, objects produce no detectable echoes. Because of the characteristics of interference patterns, the width of that cone is directly proportional to the operating wavelength and inversely proportional to the size of the array. With every element radiating in phase, the beam is in effect steered straight ahead.

Now suppose the signals from each of the radiating elements are delayed electronically by amounts that increase steadily across the face of the array. Each delay causes a signal to lag a fraction of a wavelength behind the signal from the adjacent element. A change in the relative phases of the signals is the result. Now the zone in which the individual signals add up in phase to produce a strong sum signal, capable of detecting targets, lies not straight ahead, down the boresight of the antenna, but off to the side in the direction of increasing phase delay. The angle of the beam reflects the magnitude of the phase shift, the size of the array and the wavelength of the signals. Again the beam takes the form of a slender cone surrounded by regions of destructive interference. The radar beam has been deflected without a physical movement of the antenna.

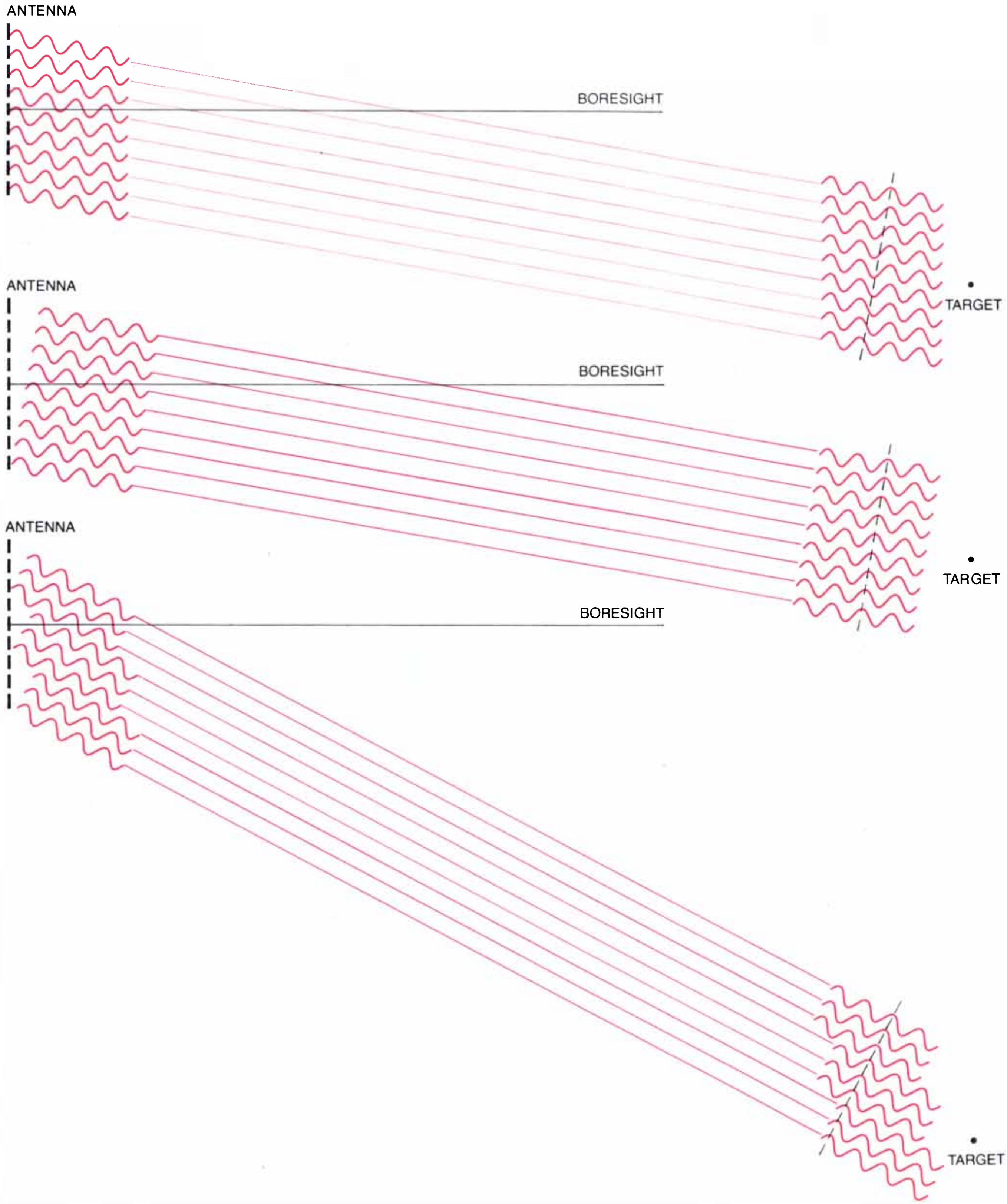
When an echo returns from a target that lies along this new, electronic boresight created by the progressive phase delays, the circuitry that delayed the transmitted signal introduces a new set of delays into the separate signals received at each of the radiating elements. As the returning wave front sweeps obliquely across the face of the array, the elements that transmitted last—those nearer the target—receive the reflected pulse first. Therefore the same set of delays that steered the transmitted pulse brings all the echo's constituent signals back into phase, ready for the processing that extracts information about the target.

Phase-lag steering enables a typical array to deflect its beam by as much as 60 degrees from its perpendicular axis, giving it a field of view 120 degrees across: a third of the way around the horizon and, if the array face is tilted sufficiently, from the horizon to well past the zenith. Since the steering entails no mechanical readjustments, moving the beam through its full range takes only a few microseconds. With a computer calculating the proper phase shifts for the desired beam angles and controlling the delaying circuitry, a phased array such as the COBRA DANE can track several hundred targets simultaneously.

The electronic hardware crucial to phase-lag steering, the device that creates the appropriate delay in the microwave signal fed to each element, is known as a phase shifter. The phase shifter consists of precisely sized cables or waveguides and works in the following manner. Increasing the length of cable through which a signal must travel on its way from oscillator or amplifier to radiating element introduces a delay in the transmission of the signal. In practice it is unworkable to allow the length of cable feeding each radiating element in a phased array to



ANTENNA ELEMENTS IN A PHASED ARRAY emit separate microwave signals. When all the elements radiate precisely in phase, yielding wave crests that move forward in step, the waves become superposed along the perpendicular axis of the array. They interfere constructively to produce a strong sum signal, resulting in a beam directed straight ahead.



ELECTRONIC BEAM STEERING requires progressive shifts in signal phase. Radiating in phase, every antenna element emits a signal that propagates in all directions. Away from the array's boresight, or perpendicular axis, the crests and troughs of the separate signals do not coincide (*top*). In this direction the individual signals are out of phase; the radar pulse is therefore weakened through destructive interference and will not produce a detectable echo from an object. To shift the region of constructive interference toward the object, in effect steering the beam, the individual signals are delayed by fractions of a wavelength. The result is a series of phase shifts that increase across the face of the array. The phase-steered signals now coincide at an angle to the boresight (*middle*), interfering constructively to produce a strong signal. To deflect the beam further the increment between successive phase shifts must increase. To reduce the necessary circuitry the delays never exceed a full wavelength; instead steep increases from zero delay to the maximum phase shift of about one wavelength occur repeatedly across the face of the array (*bottom*). The result is a beam in which the crests of individual signals coincide at an extreme angle to the boresight, resulting in constructive interference and a strong signal.

be varied infinitely, yielding continuously variable phase delays. Instead the phase shifting proceeds in steps. A hierarchy of cables, each a different length, is attached to each element. A particular combination of the cables is switched into each circuit to achieve the phase shifts necessary for the desired deflection of the beam.

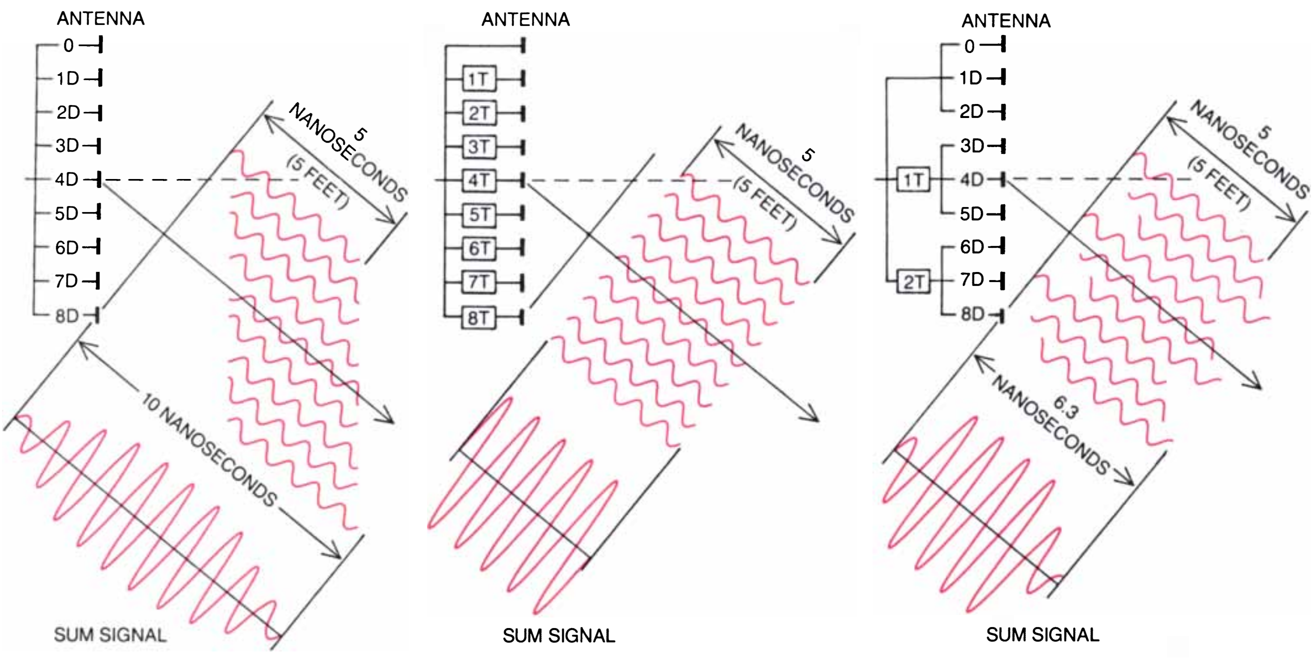
The COBRA DANE intelligence radar, for example, uses three-bit phase shifters: phase shifters with three path lengths. Each shifter contains three lengths of stripline, a kind of waveguide that channels microwave energy along a narrow copper strip centered between two copper ground planes. One of the striplines adds half a wavelength to the path of the signal, about six inches, at the COBRA DANE's operating frequency of about a gigahertz (one billion cycles per second). It thus introduces a phase shift of 180 degrees with respect to an undelayed signal. Another stripline delays the signal by a fourth of a wavelength, for a phase shift of 90 degrees, and the third is equivalent to an eighth of a wavelength, or a 45-degree phase shift. In combination the three striplines can change the phase of a signal by any multiple of 45 degrees, from zero to 315 degrees.

Stepped phase delays would seem to imply blind spots. How can a mere eight possible phase delays, at intervals of 45 degrees, result in a continuously steerable radar beam? The answer is inherent in the characteristics of interference patterns. Each time the difference in phase between the signals radiated from opposite sides of the array is increased by 360 degrees, or one wavelength, the region of constructive interference that makes up the beam will move by approximately its own width. To shift the perpendicular beam that results when all the signals are in phase to a contiguous position, leaving no intervening blind spot, the total phase shift across the face of the array must be about 360 degrees.

Whether the individual phase shifts increase steadily across the array face or in steps of 45 degrees is unimportant; coarser increments merely result in a slight drop in beam strength and in antenna sensitivity. To achieve finer steering an antenna equipped with three-bit phase shifters can make use of smaller total phase changes: 180 degrees, for example, in four steps of 45 degrees.

If the beam is to be deflected by more than its width from the perpendicular, the total phase change across the array face must increase beyond 360 degrees. Because of the periodic character of electromagnetic waves, a phase shift of multiple wavelengths is equivalent to a 360-degree shift. For a total phase change of more than 360 degrees the linear increase in phase delay from zero to 360 degrees must be repeated a number of times across the face of the array. The first series of phase lags shifts the phase by a total of one wavelength; the second increases it from one wavelength to two wavelengths, and so on. A graph of phase lag versus distance across the array face would resemble a series of sawteeth; the steeper their sloped sides are and the greater their number is, the sharper the beam deflection will be.

Simple geometry shows that the effective area of the antenna decreases with increasing steering angle. As a result the sensitivity of a phased array to echoes returning from a target drops rapidly at angles beyond about 60 degrees away from the array's perpendicular axis. By itself a phased array cannot match the ability of a mechanically turned dish to scan in all directions. One solution is to link a number of arrays facing in different directions. Another means of expanding the coverage of a phased array is to mount the array horizontally under a domed lens that refracts the beam, an arrangement that increases its angular deflection. When the array produces a beam aimed 60 degrees from the zenith, the lens might bend it further until it trav-



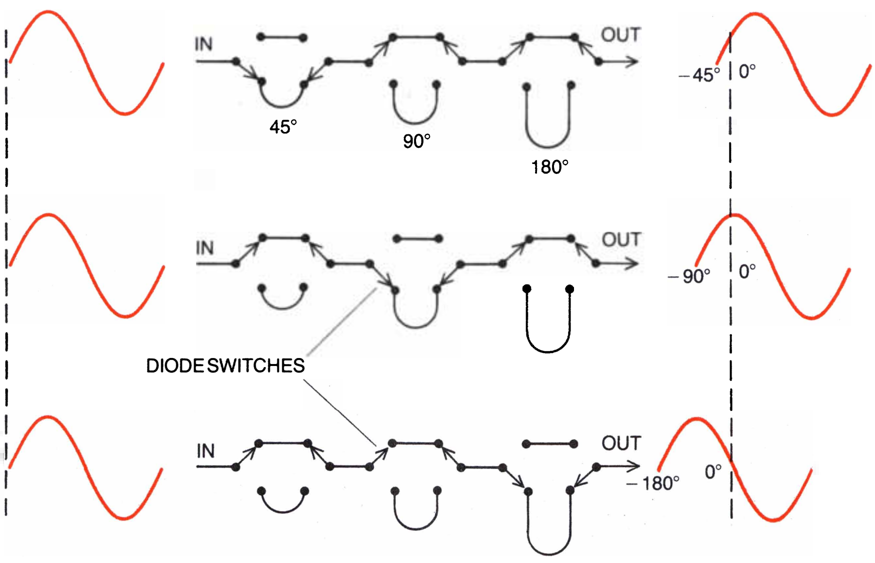
TIME DELAYS keep a very brief radar pulse, needed to determine the size of an object, from being stretched and distorted when it is transmitted at a sharp angle to the boresight of the array. A five-nanosecond (five billionths of a second) pulse measures about five feet from beginning to end when it is aimed straight ahead, but when it is deflected sharply through phase shifts, represented by multiples of *D* (the delays in the individual waveforms are omitted for clarity), it may become stretched to several times that length (*left*). Progressive delays, represented by multiples of *T* and amounting to many times the lags used for phase steering, preserve the pulse's shape and dimensions (*middle*). To avoid the complexity of separate time-delaying circuitry for each of hundreds or thousands of antenna elements, many phased arrays are divided into subarrays, each one steered through phase shifting. At sharp steering angles the subarrays produce distorted signals, but they are coordinated through time delays so that the distortion is not compounded across the face of the array. The dimensions of the pulse remain fairly accurate and fewer time-delaying circuits are needed (*right*).

els at right angles, straight toward the horizon. The lens thus gives the array the capacity to survey the entire hemisphere of sky. Such a lens can consist of a specialized ceramic or plastic that refracts microwaves; it can also take the form of a second set of phase shifters, preset to introduce additional phase lags into the signal they receive from the array.

When phase steering is used to direct a brief pulse of radar energy at a large angle with respect to the boresight of the array, the emerging pulse is inevitably distorted: stretched in time and space. Assume an antenna transmits a pulse lasting for five nanoseconds (five billionths of a second). If the radar is steered straight ahead, the pulse that leaves the antenna has a rectangular cross section, its breadth equal to the width of the array and its length equal to the distance electromagnetic waves travel in five nanoseconds, about five feet. If, on the other hand, phase shifting has sharply deflected the beam from the boresight of the array, the pulse will take the shape of a parallelogram. From the perspective of a target the length of the pulse will be greater than five feet since its component signals, instead of striking simultaneously, will reach the target in a sequence. The echo that returns will also be stretched.

For detecting and tracking targets a radar ordinarily transmits much longer pulses—1,000 nanoseconds long, for example—and a few nanoseconds' distortion is of little consequence. The stretching has little effect on the capacity of the radar to glean information about position and velocity from the echo. A brief pulse must be transmitted, however, to resolve objects traveling in close formation. A brief pulse is also crucial in radars designed to gauge the size of a target from the distinct reflections that return from its fore and aft surfaces when it is illuminated by an extremely short pulse. If a brief transmitted pulse is stretched sufficiently, echoes that might otherwise be distinct can merge, obscuring such information.

A procedure similar to phase-lag steering overcomes this difficulty by preserving the shape of the pulse. In phase steering it is necessary to delay signals by only fractions of a wavelength; the delays necessary to avoid pulse stretching are equivalent to multiple wavelengths. In effect the individual signals are given head starts proportional to their distance from the target; the effect is the same as if the array were rotated to place the target on the boresight. The process is known as time-delay steering; like incremental phase lags, it yields a signal at the desired steering angle that is coherent and therefore strong.



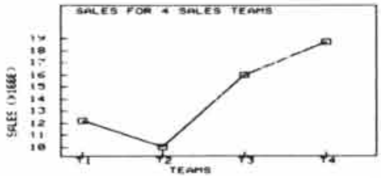
PHASE SHIFTER routes the microwave signal that is supplied to each antenna element through cables of varying length. The cables delay the signal, thereby shifting the relative phase of the output. Because microwaves, like all electromagnetic radiation, travel at a fixed speed, the delay each pathway creates is proportional to its length. The illustration shows the three basic delays each phase shifter can introduce. The shortest of its three pathways results in a delay of 45 degrees, or an eighth of a wavelength (*top*). If that pathway is bypassed and the next pathway is switched in, the phase shift is 90 degrees, or a quarter of a wavelength (*middle*). The longest of the three pathways yields a delay of 180 degrees, or half a wavelength (*bottom*). Combinations of the pathways can produce phase shifts, in increments of 45 degrees, of up to 315 degrees. For each steering angle of the radar beam the central computer calculates the proper phase delay for each of the radiating elements and switches in the appropriate combination of phase-shifter pathways.

Such long delays, the equivalent of many feet of signal travel, require that proportionate lengths of cable be incorporated into the path the signal follows from the oscillator or the amplifier to the radiating element. A large phased array can include many thousands of radiating elements; if each of them incorporated time-delay circuitry, an enormously elaborate and expensive radar installation would result. Radar designers instead strike a compromise between accurate pulse shape, even at high deflection angles, and engineering simplicity. The result is an array steered by both phase shifting and time delays.

In the COBRA DANE, for example, each of the 15,360 radiating elements is linked to a three-bit phase shifter; hence each of the signals is phase-shifted individually. In searching for targets the radar transmits 1,000-nanosecond pulses and is steered only by means of phase lags. Since the purpose of the radar is to monitor Soviet ballistic-missile tests, however, it must gather information on size once it has located an object. To that end its antenna is organized into 96 subarrays, or blocks, of 160 radiating elements each. Once the radar has detected a target it begins to transmit extremely short pulses, and time delays are introduced into the signals fed to each subarray. The time-delay circuitry is a larger analogue of the phase shifters: it consists of coaxial cable in a range of fixed lengths, any combination of which can be switched into the circuit to achieve time delays of from one wavelength to 64 wavelengths—about 64 feet, at the COBRA DANE's operating frequency of about a gigahertz.

Because the subarrays measure only about nine feet across, in contrast to the 95-foot diameter of the complete array, the distortion that arises within each subarray at sharp steering angles is moderate. Each produces a signal in the shape of a parallelogram, but those signals are summed through time-delay steering, and so their individual distortion is not compounded. The result is reasonable preservation of the shape of the pulse, and the need to use only 96 time-delay units rather than 15,360. In terms of material, time-delay steering in the COBRA DANE requires only about a mile of additional cable rather than the 100 miles that would otherwise be needed.

The replacement of a single, movable antenna with an immobile array of radiating elements can offer more mundane advantages than elec-

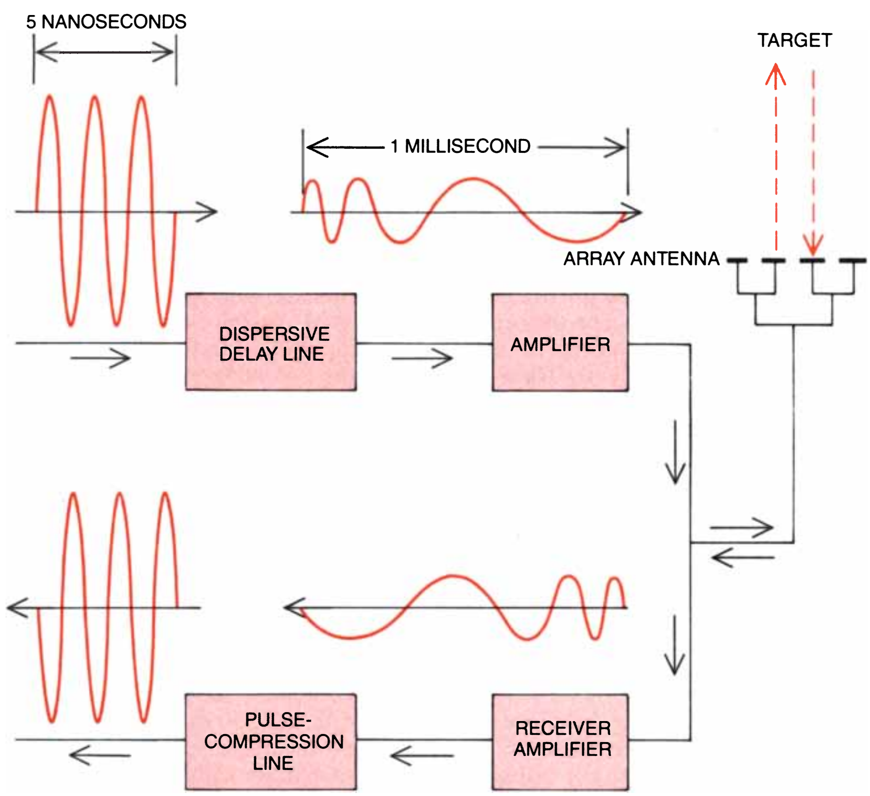


tronic steering. Reliability is one of them. A fixed array does not depend on fallible mechanical components such as bearings and motors. Furthermore, most mechanically steered radars use one or at most a few very large microwave tubes to amplify their signal. To take one example, the Marconi Martello, a British-made air-defense radar, is built around a single tube with a power output of about three megawatts. If the tube fails, the system fails with it. Such defense and intelligence radars, however, can switch quickly to auxiliary sources of power.

The COBRA DANE, in contrast, is powered by 96 tubes, each of which produces 160 kilowatts of power. The output of each tube passes through a divider and thence to the 160 radiating elements of one subarray. Tube failure in the COBRA DANE therefore incapacitates only one ninety-sixth of the array, leaving the radar as a whole still functional, at slightly reduced performance. Moreover, the smaller tubes are easier to replace than the single large tube of the Martello radar.

Solid-state phased arrays can deliver still higher levels of reliability and ease of maintenance. The radars code-named PAVE PAWS, which are designed to detect sea-launched ballistic missiles from vantages on Cape Cod and in California and from planned sites in Georgia and Texas, generate their power using transistors. Four 100-watt transistors are hooked in parallel in a separate module driving each radiating element. Thus on each of the twin PAVE PAWS antenna faces the job of amplifying the signal is divided among the 1,792 antenna elements rather than among 96 tubes, so that the effect of a single failure on the performance of the radar is smaller still. Moreover, the mean time between failures for a solid-state module is considerably longer than that for a COBRA DANE tube: 100,000 versus about 20,000 hours. When the foot-long, 28-volt modules do fail, they are much easier to replace than the COBRA DANE's five-foot-high, 40,000-volt tubes.

In the PAVE PAWS, as in many solid-state radars, the amplification takes



PULSE CODING AND COMPRESSION enable a radar to transmit a long pulse but obtain size and distance information that is as accurate as when a short pulse is used. A short pulse usually requires impractically high transmission power for adequate range. In pulse coding a very brief signal consisting of a range of frequencies passes through a dispersive delay line in which its components are delayed in proportion to their frequency. In the process the pulse is stretched; for example, a five-nanosecond pulse might be lengthened by a factor of 200,000 to a duration of one millisecond. It is then amplified and transmitted; an echo returns from the target and is itself amplified. A pulse-compression line retards the echo by amounts that vary inversely with frequency to reduce the signal to its original five-nanosecond length. The compressed echo yields all the information that would have been available if an unaltered five-nanosecond pulse had been transmitted.

place after the signal has been divided among the antenna elements and its phase has been shifted. As a result the power losses that result when an amplified signal passes through divider and phase-shifter circuitry are eliminated. For this gain in efficiency, however, and for all the other advantages of solid-state technology, there is a trade-off: in general, solid-state circuitry yields lower peak powers than those available to a tube-powered radar.

Limitations on signal power have added to the importance of techniques known as pulse coding and pulse compression, which enable a radar to simulate a brief, high-powered pulse by emitting less powerful signals for a longer period of time. That capacity is important even for high-powered tube radars, both mechanically steered dishes and phased arrays, in gleaning certain kinds of information about distant targets.

The distance at which a radar of a given sensitivity can detect an object of a particular size and reflectivity depends on the total energy of a pulse. The shorter the pulse is, the higher the peak transmission power must be if the range of the radar is to remain unchanged. The COBRA DANE can detect a metallic object the size of a grapefruit at a distance of 1,000 nautical miles; to do so with a five-nanosecond pulse it must radiate a peak power of three trillion watts—more than enough power to destroy all its circuitry.

Yet a brief pulse is essential if the radar is to gauge an object's size or distinguish a number of objects flying in close formation. The fact that range is a function not of peak power but of total pulse energy implies a solution: stretching the pulse, reducing the necessary peak power correspondingly, in a process known as pulse coding, then compressing the echo in order to extract from it all the information an actual brief pulse would have supplied. In the COBRA DANE, for example, a five-nanosecond pulse is stretched by a factor of 200,000, to a duration of one millisecond, before it is amplified and transmitted. The peak power required drops by an identical factor: from three trillion watts to 15 megawatts, the COBRA DANE's actual power.

In a typical instance of pulse coding a five-nanosecond pulse composed of a range of frequencies passes through a dispersive delay line, which retards its components by amounts that increase with frequency: the lowest frequencies in the pulse receive no delay, whereas the highest frequencies receive the maximum delay of a millisecond. Now one millisecond long, the pulse is amplified and transmitted; an echo of equal duration returns.

The echo is fed through a receiver compressor network, which adds a second set of delays. This time the delays follow an inverse relation to frequency. The lowest frequency in the pulse receives a millisecond of delay and the highest frequency no delay at all. Each of the signal's components has been subjected to the same total delay (one millisecond) through pulse coding and compression. The result is an undistorted five-nanosecond echo.

If the one-millisecond-long transmitted pulse, which measured about 186 miles in length, struck a substantially shorter object, it returns as a pair of overlapping echoes. Ordinarily such echoes cannot be separated and their relative timing cannot be used to gauge the object's size. When the pulse-coded echo is compressed, however, the overlapping echoes resolve themselves into two distinct five-nanosecond signals.

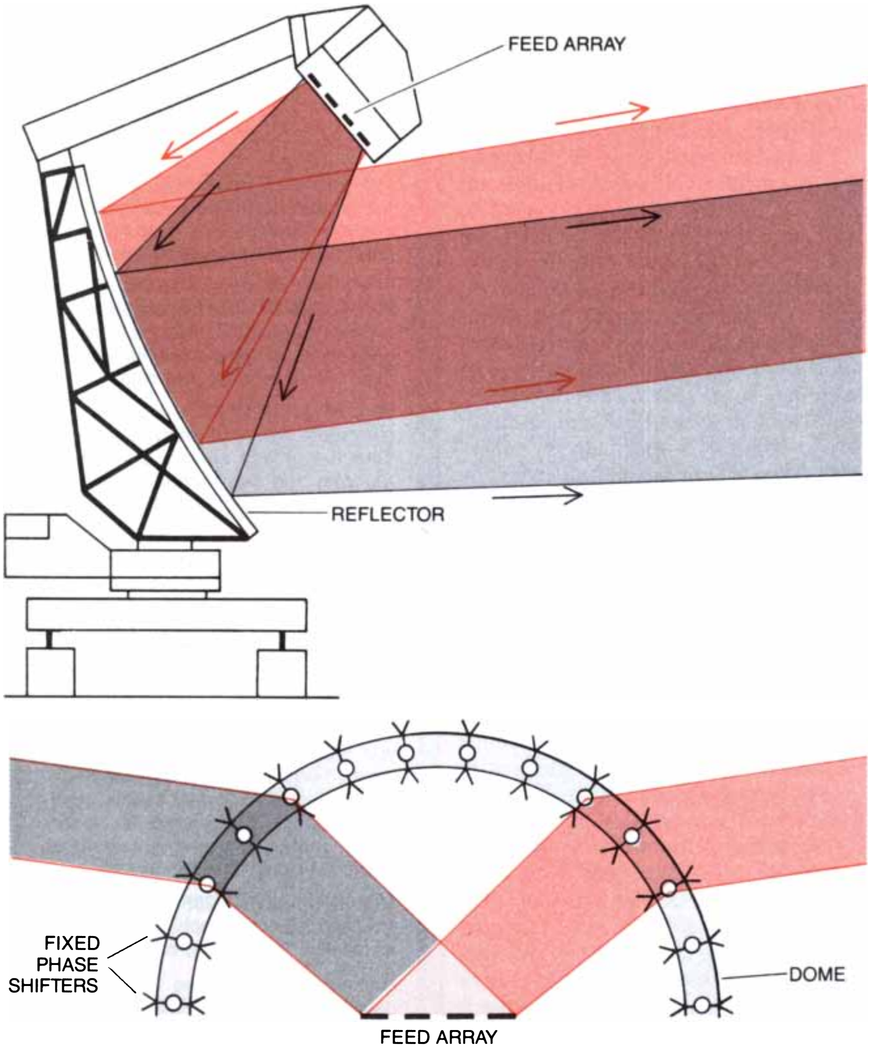
Pulse coding and pulse compression play a parallel role in solid-state radars. Even when there is no need to determine the size of the object from which the radar signal reflects, an accurate determination of distance requires a reasonably short pulse. Without pulse compression a one-millisecond pulse can resolve the distance of an object only to within 93 miles. Moreover, long pulses are susceptible to "clutter"—extraneous reflections from precipitation and from the ground. Yet the high powers needed for adequate range at shorter pulse durations are beyond the reach of many solid-state radars. Pulse coding and pulse compression enable such radars to reconcile range and resolution with low power.

The first phased-array radars to be installed, dating from the 1960's and 1970's, serve military and intelligence purposes. There are circumstances in the civilian world in which the same requirement arises that spurred the development of the military technology: the need for frequent data on a number of fast-moving objects. One example occurs near the end of airport runways, where incoming aircraft line up for their final approach. There precision-approach radars transmit pulses to incoming planes, simultaneously tracking their distance and position in relation to the end of the runway and guiding them down. Increasingly, such radars make use of the technology of phased arrays.

With smaller numbers of radiating elements the cost of a phased array declines. For most uses the number of radiating elements is necessarily high. A small array transmits a beam that is less focused and hence wider. This factor reduces the angular resolution of the array, and the small antenna area reduces the array's sensitivity to echoes. When it is not necessary that a broad area of sky be surveyed, however, both drawbacks of a small array can be surmounted by combining the array with a large reflector.

The field of view of a precision-approach radar need not be large: typically it must scan about 10 degrees of azimuth and from seven to 14 degrees of elevation. As a result a hybrid of phased array and traditional reflector works well. In one design an array of 443 radiating elements is combined with a reflecting dish measuring 15 by 13 feet. The array is positioned near the focus of the dish, which reflects the beam at every steering angle. In the process the reflector acts as a lens, focusing the beam while reducing its angular deflection. The reflected beams are narrower and are confined within a narrower angle of space; as a result the ability of the array to resolve two targets lying within a small angle, as well as its ability to determine the precise bearing of a single target, is improved. The dish also increases the radar's sensitivity to echoes.

Future phased-array radars will reflect advances in circuit technology. The introduction into radar engineering of the equivalent of digital integrated circuits will considerably reduce the number and size of the components needed to generate, receive and process the signal. Known as monolithic microwave integrated circuitry (MMIC), this new technology unites phase shifters, switches and transistorized amplifiers on gallium arsenide chips. A complete transceiver (transmitter/receiver) module, containing all the circuitry needed for a single radiating element in a phased array, can already be assembled from only 11 MMIC chips. In contrast, the solid-state transceiver modules currently in use require hundreds of parts.



HYBRID RADARS combine a phased array with a lens that alters the characteristics of the beam. In a limited-scan phased array (*top*), a design increasingly used for airport approach radars, a small phased array, which produces a relatively broad beam, emits its energy into a hyperbolic reflector. The beam is pictured at two steering angles to show how the reflector focuses it, reducing its dispersion while also reducing its angular deflection. The reflector thus increases the angular resolution of the array but narrows its field of view. The arrangement allows a smaller, cheaper array to be used in circumstances that do not demand a broad scanning angle. A dome-shaped lens that covers a horizontal array in an experimental military radar (*bottom*) has the opposite effect, increasing the deflection of the beam. The lens consists of either a specialized ceramic or plastic material that refracts microwaves or a domed configuration of phase shifters that add fixed increments of phase delay to the beam they receive from the feed array, in effect resteering it at a sharper angle. Such a radar can electronically scan from the horizon to the zenith in every direction.

Ultimately such a module may also include a microprocessor that will control the phase shifting, calculating the proper phase delay when a central computer specifies the elevation and azimuth of the beam. In present phased arrays that function devolves on the central computer, which sequentially calculates the phase changes for the thousands of radiating elements. Phase-change calculation and control at individual elements would eliminate the thousands of cables currently used to carry the phase-shift commands from the computer to individual antenna elements. With advances in circuit technology the process of phase shifting itself may eventually become a digital process, done as a mathematical operation on a signal in digital form. Such an advance would allow the width of the beam to be changed at will: broadened for the scanning of broad areas of sky or focused for precise tracking. Meanwhile similar digital processing of the echo would aid in the elimination of interference from a jamming signal.

Increasing simplicity and compactness, with concomitant reductions in weight and increases in reliability, will soon make phased arrays suitable for duty in space. Orbiting phased arrays might serve to map the earth's surface, monitor shipping traffic and icebergs and perform military and surveillance duties. The National Aeronautics and Space Administration's next generation of mapping radars, to be carried on board the space shuttle, will be phased arrays, and both the Navy and the Air Force are considering launching a satellite-borne phased array. Many of the radars launched to date could not be steered; they surveyed only a narrow strip in the course of each orbit. The swift, unconstrained electronic steering of a phased array, in contrast, would put vast swaths of the earth's surface within the compass of its beam.


Photo: © Alan Bergman

The Functional Replacement of the Ear

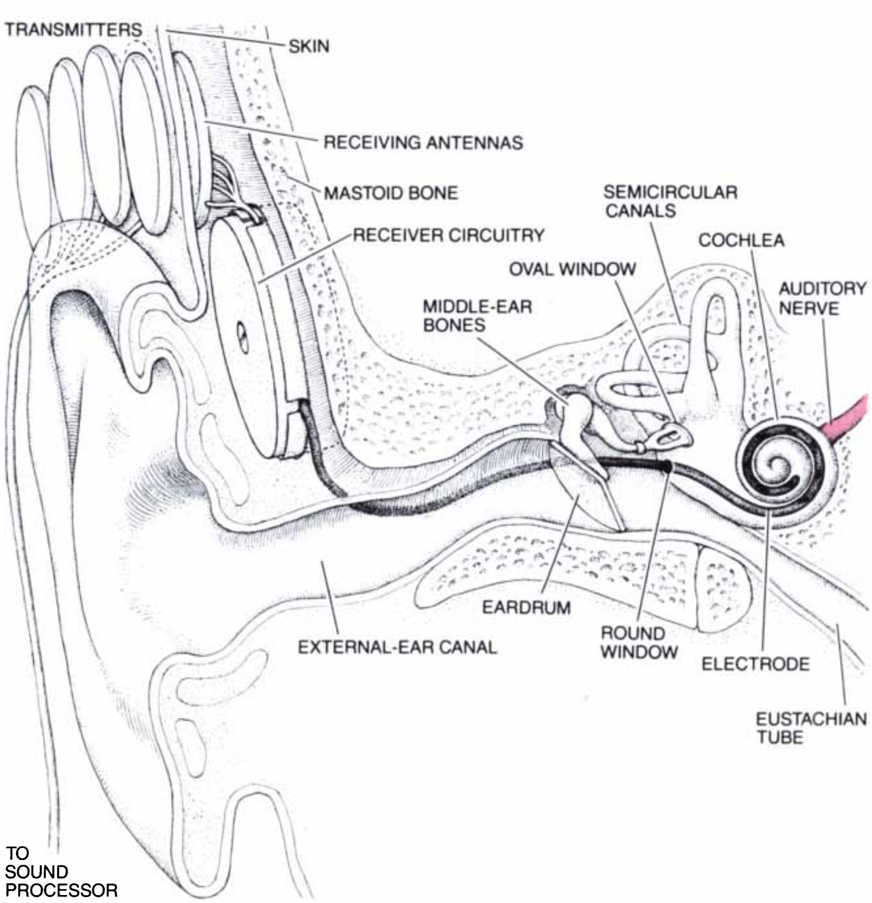
Implantable prostheses designed to deliver electrical stimuli directly to the auditory nerve hold considerable promise for people with a type of deafness in which the sensory hair cells of the inner ear are damaged

by Gerald E. Loeb

Recent advances in the fields of electronics and neurophysiology have led to the emergence of the new field of neural control. This experimental discipline relies on the exchange of information between an electronic circuit and a nervous system for the purpose of studying or supplementing a biological function. One of the main objectives of such research is the development of prosthetic devices for replacing defective parts of the human nervous system. Much of the progress in this area has resulted from long-term collaboration between teams of investigators engaged in both basic and applied research under the aegis of the Neural Prosthesis Program of the National Institute of Neurological and Communicative Disorders and Stroke. My own work on the functional replacement of the human ear has been done as part of such a project with a team at the University of California at San Francisco.

The cochlear implant, the particular neural prosthesis that is the subject of this article, is intended for patients with sensorineural deafness. In such patients the functioning of the sensory hair cells of the cochlea, the snail-shaped structure at the core of the inner ear, is impaired. In normal hearing sound travels through the external ear canal to the tympanic membrane, or eardrum, which transmits the vibrations in the air to a system of small bones in the middle ear. The innermost of these bones, called the stirrup, contacts the oval window, a membrane-covered opening in the base of the cochlea, relaying the vibration to the fluid-filled interior of the cochlea, where it is sensed by the hair cells in a structure known as the organ of Corti. The hair cells are arranged in four long rows on the basilar membrane, a flexible partition at the bottom of the organ of Corti that separates two of the cochlea's three parallel, spiral canals. The hair cells convert the vibration of the basilar membrane into an electrical signal, which travels along the auditory nerve to the brain.

Several kinds of cochlear implant have been designed to circumvent this elaborate transmission process in patients with sensorineural deafness. All the devices have four features in common: a microphone for picking up the sound, a microelectronic processor for converting the sound into electrical



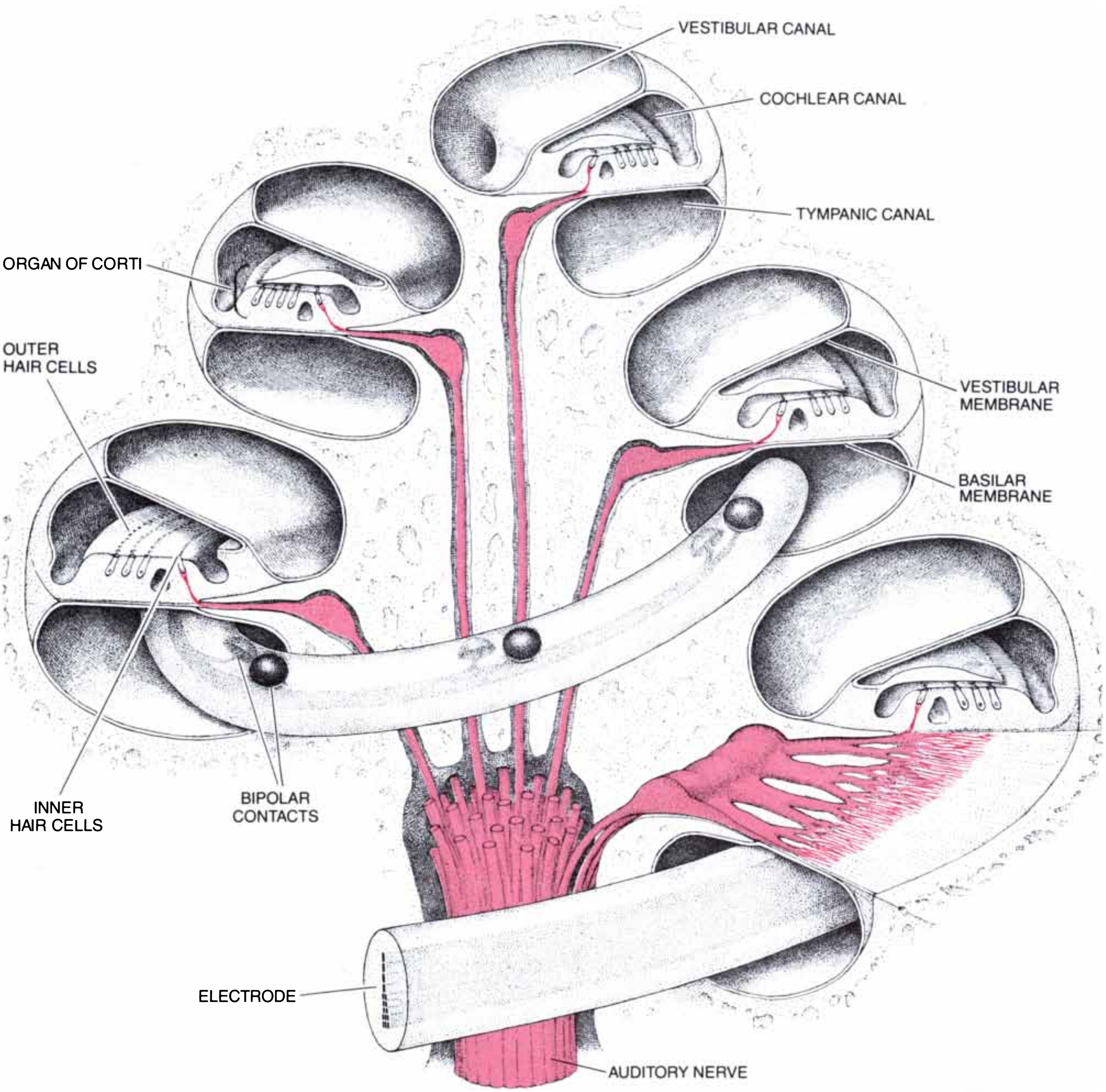
EXPERIMENTAL HEARING AID developed by workers at the University of California at San Francisco is shown implanted in a patient's ear in the cutaway drawing at the left. An enlarged interior view of the cochlea, the snail-shaped structure at the core of the inner ear, appears at the right. (The rows of sensory hair cells, included in this view, would be absent in a typical patient with sensorineural deafness.) The particular cochlear implant depicted is an eight-channel, bipolar device: it delivers stimuli at eight different frequencies

signals, a transmission system for relaying the signals to the implanted components, and a long, slender electrode that a surgeon snakes into the inner recesses of the cochlea so that the device delivers the electrical stimuli directly to the fibers of the auditory nerve in one or more places.

One such device, a comparatively primitive single-channel model, has recently been approved by the Food and Drug Administration for implantation in patients with sensorineural deafness. Developed by William House of the House Ear Institute of Los Angeles, it is manufactured by 3M [see "Tuning a Deaf Ear," "Science and the Citizen"; SCIENTIFIC AMERICAN, November, 1984].

More sophisticated single- and multichannel devices, currently in the research stage, promise to provide more realistic sound perception to such patients. The intensive study of the responses of these patients and of experimental animals to various patterns of electrical stimulation from such devices has the added benefit of providing fresh insights into the normal encoding and decoding of acoustic information by the nervous system.

In one sense the structure of acoustic information is very simple. The signal picked up by a microphone or an ear can be fully described by a single function of time, which characterizes the motion of a single point in space, such as the center of the microphone's diaphragm or the cochlea's oval window. In spite of its simplicity, such a signal typically contains complex



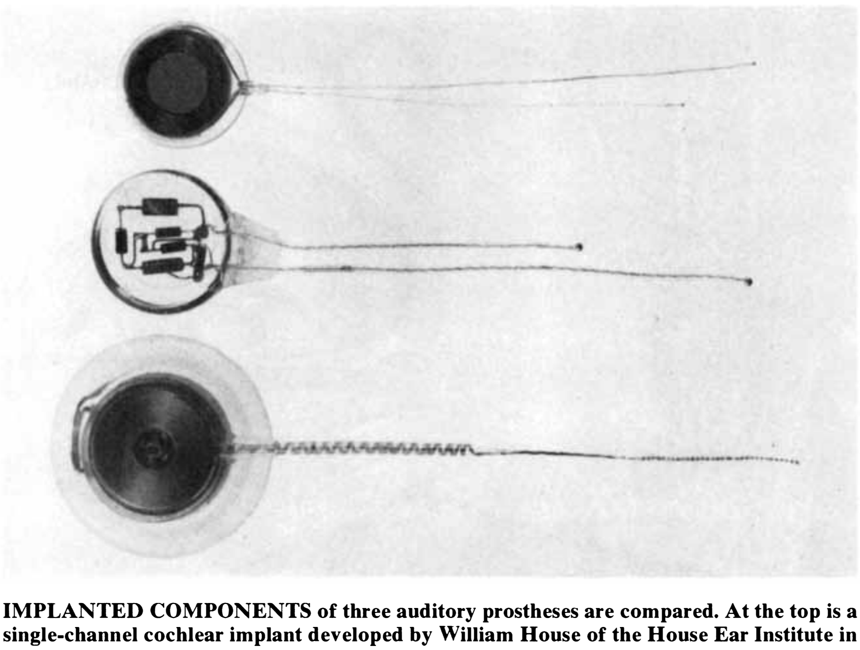
to separate groups of auditory-nerve fibers in the cochlea by means of eight closely spaced pairs of electrical contacts distributed along the length of the implanted electrode. Multichannel devices of this kind, currently in the clinical testing stage, are expected to provide more realistic sound perception to patients with sensorineural deafness than the single-channel devices now available. The operation of the main components of a four-channel "driver" for this device is shown in the illustration on page 108. Although the present driver has only four channels, the electrode (which may not be readily replaceable because of scar tissue) has been designed to accommodate eight channels; a surgically accessible connector in the base of the driver facilitates its replacement with improved versions.

information about the source of the sound; the nervous system is able to analyze the signal to extract the information. In the case of a human speaker the analysis is usually sufficient to identify the particular speaker and his words. The addition of a second information channel at a separate point in space—the other ear—makes it possible to distinguish multiple sound sources by their relative positions.

In the course of this analysis the one-dimensional function must be broken down into its component frequencies. The separation of such a complex waveform into its spectral components by means of the time-honored mathematical method of Fourier analysis is now a common practice in electronics and is often available in modern test instruments at the push of a button. Since the late 19th century it has been assumed that the brain must be employing a comparable form of signal analysis in at least the first stages of hearing. It has also been recognized that the channels through which information flows in the nervous system, the individual neurons, are inherently very slow. Somehow an analytic system built up of a large number of information channels, each channel limited to about 300 pulses per second, must provide an accurate and almost instantaneous spectral analysis of signals covering a bandwidth of between 20 and 20,000 hertz (cycles per second). Furthermore, in spite of the presence of considerable noise in each channel, the overall performance of the system must not be seriously degraded over a dynamic range of a million to one (120 decibels) from the threshold of hearing to the beginning of pain.

How does the ear accomplish this demanding task? Hermann von Helmholtz was one of several 19th-century physicists who recognized that the organ of Corti could be involved in some way in sampling sounds that were separated physically into their spectral components by the sympathetic vibration of the basilar membrane. Pioneering studies of the motion of the basilar membrane by Georg von Békésy in the 1950's established that when sound of a given frequency is applied to the base of the cochlea, it causes the basilar membrane to vibrate with the largest amplitude at a particular place, which is mechanically tuned to the frequency of the applied sound. A complex acoustic waveform, a composite of many individual sinusoidal frequencies, generates a spatial distribution of sympathetic vibration peaks along the basilar membrane; the higher frequencies are represented at the base of the cochlear spiral and the lower frequencies at the apex. The hair cells of the organ of Corti transduce the mechanical motion of the basilar membrane into electrical activity in the adjacent auditory-nerve fibers. As a result low-frequency impulses are generated in the parallel array of auditory-nerve fibers. Each fiber transmits information only about the amplitude of the vibration in a particular place.

This "place pitch" theory of sound perception underlies the design of most multichannel cochlear prostheses, although it will become apparent from what follows that the place-pitch mechanism is probably not sufficient to account for many important psychophysical phenomena of hearing, including some critical attributes of normal speech perception.



IMPLANTED COMPONENTS of three auditory prostheses are compared. At the top is a single-channel cochlear implant developed by William House of the House Ear Institute in Los Angeles. Manufactured by 3M, it is the first such device to receive the approval of the Food and Drug Administration. The implanted components consist of a simple wire coil leading directly to a pair of ball-shaped electrodes, one of which is inserted in the tympanic canal at the base of the cochlea. (The other electrode is grounded to a nearby part of the middle ear.) In the middle is a more advanced single-channel device developed by Inge J. Hochmair-Desoyer and Erwin S. Hochmair of the Technical University of Vienna and currently being tested by 3M. The receiver circuitry can be seen inside the clear plastic enclosure; in this case both ball electrodes are designed to be placed outside the cochlea in the middle ear. At the bottom is a multichannel device developed by Graeme M. Clark and his colleagues at the University of Melbourne and manufactured by Nucleus Limited. The hermetically sealed titanium enclosure of this device contains the complex circuitry needed to decode a transmitted signal that selects and actuates one of 22 contacts along the electrode.

A cochlear implant relies on the fact that many of the auditory-nerve fibers often remain intact in patients with sensorineural deafness. As in the case of most electrically excitable cells, the surviving neurons can be stimulated to fire actively propagating nerve impulses by applying external electric currents of the proper strength, duration and orientation. Such "evoked potentials" arrive at the brain looking just like the impulses generated by acoustic signals that intact hair cells transduce; accordingly the brain interprets them as sound.

The loudness of the sound perceived depends roughly on the number of nerve fibers activated and their rates of firing. Both variables are functions of the amplitude of the stimulus current. The pitch is related to the place on the basilar membrane from which those nerve fibers once derived their acoustic input, in agreement with the place-pitch theory. In principle, with enough independent channels of stimulation, each controlling the activity of a small, local subset of the auditory-nerve fibers, one could re-create the normal neural response to acoustic stimuli of any spectral composition. The brain would then process that information in its usual manner and the subject would "hear" the "sounds."

The problem faced by the designer of a neural prosthesis is that all the auditory-nerve fibers are swimming in the same pool of electrically conducting tissues and fluids. An electric current injected into this medium tends to spread out symmetrically from the source; as a result the current density decreases as the square of the distance from a monopolar source. For such a stimulus to selectively activate a particular subset of the auditory-nerve fibers the electrode must be much closer to those fibers than to other fibers.

This constraint is a difficult one. The auditory-nerve fibers spread out to their greatest extent as they enter the organ of Corti on the basilar membrane. A longitudinal array of electrodes can be inserted in the tympanic canal and passed along the basilar membrane for a distance of about 25 millimeters from a second opening in the base of the cochlea, called the round window. Distributed along the path are the fibers that normally transmit information about sound frequencies above 500 hertz. The shortest distance from the auditory-nerve fibers to the best location in the tympanic canal (along the medial wall) is on the order of one millimeter. A pulse of stimulation current four times the threshold for fibers at one location (characterized as having a moderate loudness) will start to influence fibers two millimeters away in both directions.

When natural variations in the size of the fibers and their susceptibility to stimulation are factored in, the stimulus spread is probably even greater. Obviously only a small number of independent stimulation sites with any appreciable dynamic range could be accommodated along the available length. For speech perception the critical bandwidth is between 500 and 3,000 hertz, representing a distance along the basilar membrane of less than 14 millimeters; this distance is capable in turn of accommodating perhaps two or three independent stimulation sites that can contain a reasonable dynamic range. Experiments in which speech is simulated by means of a small number of amplitude-modulated single-tone generators, on the other hand, suggest that at least six channels must be available for intelligibility.

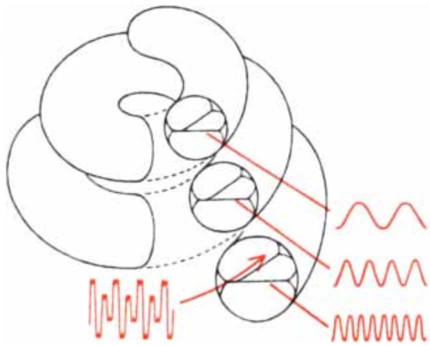
One way to effectively divide the fibers into an adequate number of discrete channels is to apply bipolar stimulation. In this approach the source and the sink of the current pulse are placed close to each other. (In monopolar stimulation the current sink needed to complete the circuit is a large, remote contact that acts as a ground for all sources.) The current-density lines around a pair of bipolar electrodes look like the magnetic-flux lines around a bar magnet: they have an elliptical shape whose major axis is oriented parallel to the line between the poles [*see illustration on page 109*].

Since the activation of neurons requires the induction of longitudinal currents along their processes, bipolar electrodes oriented at right angles to the long axis of the tympanic canal (radially to the spiral) can selectively activate the local nerve fibers passing immediately over them. Furthermore, the gradient of current density in the tissues that extends outward from a pair of bipolar electrodes is much steeper than it is for a monopolar electrode, reducing the stimulus spread even at high intensities. Tests in both animals and patients with a cochlear prosthesis have shown that such radial bipolar electrodes produce a more localized activation over wider dynamic ranges than monopolar or longitudinal bipolar configurations do.

The actual number of independent channels that can be established and their dynamic ranges depend on the condition of the remaining auditory-nerve fibers. These sometimes die back some distance from the organ of Corti, reducing their proximity to an electrode array in the tympanic canal. In some patients, however, it appears that radial bipolar pairs can be positioned about two millimeters apart without significant interaction. This spacing makes it possible for eight independent channels to span the speech-frequency sites lying between 10 and 24 millimeters from the round window.

There are several technical problems involved in actually delivering these eight bipolar stimulation channels to the human auditory nerve. The first 24 millimeters of the basilar membrane are coiled into one and a half of the two and a half turns of the cochlear spiral. The round window into the tympanic canal lies at the back of the middle-ear cavity, forcing the surgeon to work through the long, narrow external-ear canal. This means the electrode must be held in a straightened configuration as it is fed into the cochlea. It cannot, however, simply be pushed into the cochlear spiral as if it were a plumber's snake passing through a curved drainpipe. The half-cylindrical cross section of the tympanic canal causes a flexible object pressing against the side wall to be deflected upward against the extremely fragile basilar membrane. The fluid in the cochlear canal just above the membrane has a high potassium content, making it toxic to the auditory-nerve fibers in the event that it leaks into the tympanic canal.

One solution has been to use a very slender, flaccid electrode, which ideally can be coaxed into position with a minimum of trauma. Nevertheless, it is difficult, if not impossible, to control the exact orientation of the stimulation contacts of such an electrode; it tends to lie against the outer wall, far from the nerve fibers that need to be activated. The selectivity and dynamic range of the radial bipolar-electrode array depend on accurately positioning each pair on the medial wall.



PLACE-PITCH THEORY explains how the cochlear sound-detection system separates a complex acoustic waveform into its spectral components. In this representation of the theory a composite sound wave propagating in the cochlear fluid causes a sympathetic vibration of the basilar membrane, the thin, flexible partition separating two of the cochlea's three spiral canals. The vibration peaks are distributed spatially along the membrane; as shown here, the higher frequencies are detected at the base of the cochlea and the lower frequencies at the apex.

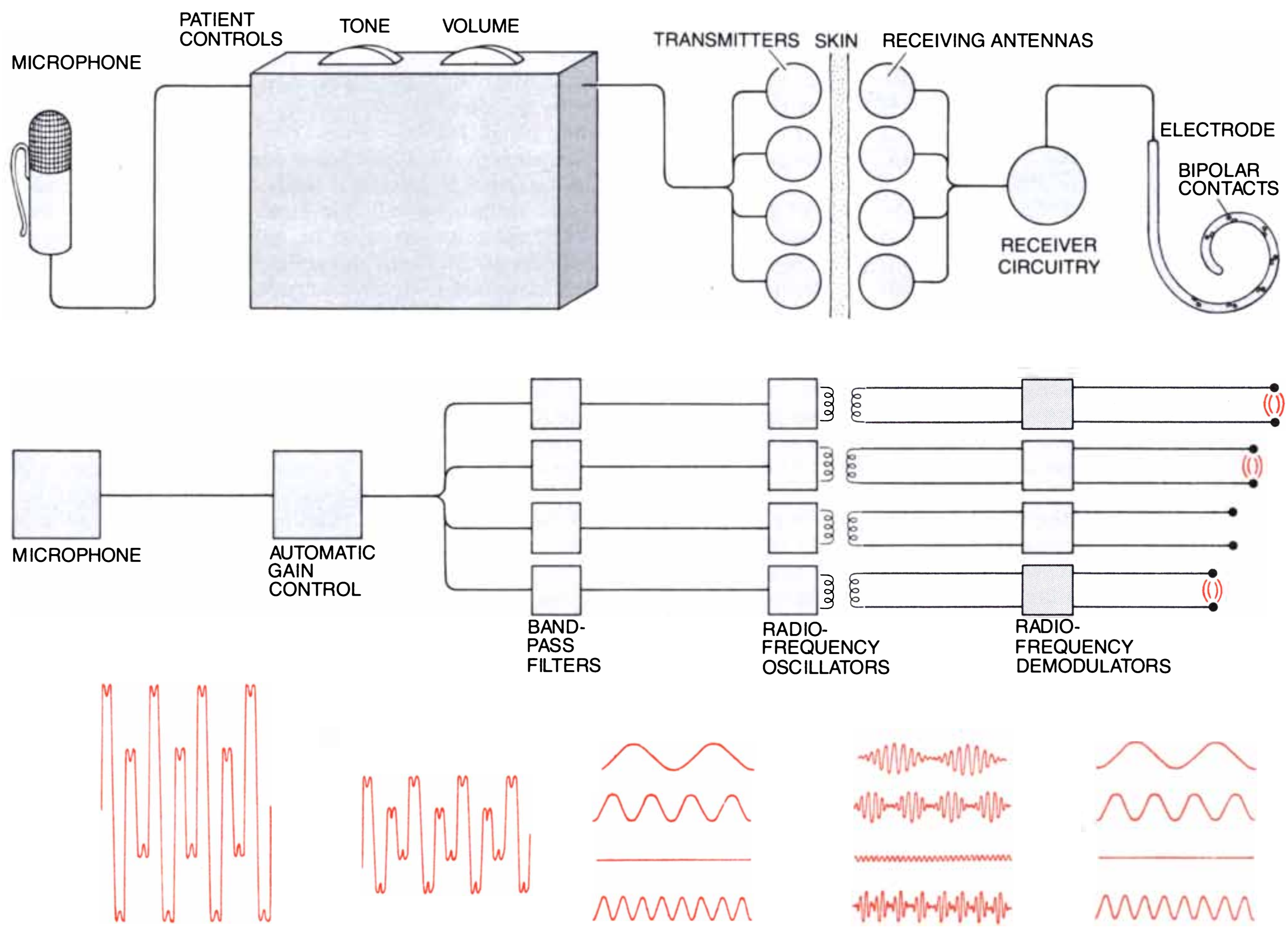
This task can be achieved only by using a comparatively thick electrode with a "memory" of a spiral shape that causes the electrode to hug the medial wall around the turns. Such handling properties have been achieved in an eight-bipolar-pair array designed by the research team at the University of California at San Francisco. It combines the mechanical properties of a silicone-rubber carrier and specially flattened and stacked platinum-iridium lead-out wires from the 16 electrode contacts distributed along the spiral form [*see illustration on pages 104 and 105*].

The delivery of stimulation current from electronic circuits to biological tissues is also a delicate matter. Electric currents in metal conductors are carried by electrons, whereas in the aqueous body fluids they are carried by ions. The electrochemical reactions at the interface of even the most biocompatible metals and the complex soup of the body's fluids is fraught with danger for both the electrodes and the tissues. Electrolytic corrosion of many supposedly inert metals is rapidly enhanced by the chelating action of chloride ions, the major negatively charged ions in extracellular fluids. This process changes the electrical properties of the interface and releases highly toxic heavy-metal ions into the immediate vicinity of the neurons, which are among the most sensitive cells of the body. The electrolysis of water alone can produce mechanically disruptive bubbles of hydrogen and oxygen, leaving behind local concentrations of hydroxyl and hydronium ions that can reach toxic levels of alkalinity and acidity respectively.

It is only in the past few years that these reactions and their consequences have been studied on a quantitative level, largely by S. Barry Brummer and his colleagues at the EIC Laboratories, Inc., in Norwood, Mass. These studies have established certain rather strict conditions for which electric currents passing through certain metals such as platinum and iridium can safely induce ionic currents in body fluids by reversible, nontoxic electrochemical reactions.

Once the electrodes are in place and the stimulation current is safely delivered, what does the patient hear? The results to date have been a generally encouraging but somewhat confusing mixture of the expected and the unexpected.

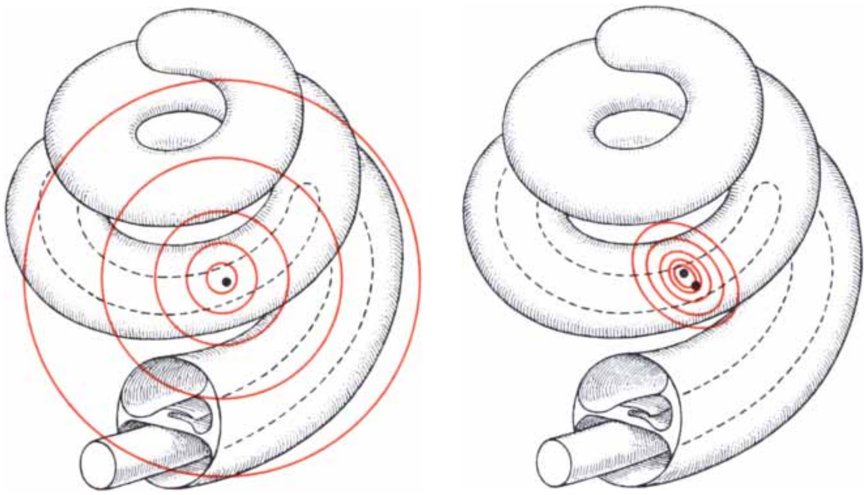
By far the largest sample of patients have been tested with single-channel stimulators. Moderately filtered waveforms taken directly from a microphone are transmitted to an implanted stimulator, which activates some auditory-nerve fibers near the base of the cochlea. Predictably, the sensations are those of a complex, amplitude-modulated noise. There is only the broadest sense of pitch and only at stimulation frequencies below a few hundred hertz. Nevertheless, a sense of the rhythm and the loudness of the sounds heard is strong. In patients who are profoundly deaf, for whom even the most powerful conventional hearing aid is no help, such sound information can provide quite helpful cues for everyday living. The lowest frequencies of the acoustic spectrum include important information about the presence and nature of ambient noises such as the sounds of telephones and automobiles, and about the loudness and cadence of speech, both for the subject's own vocalizations and for those of a person he or she may be trying to



OPERATION OF A FOUR-CHANNEL IMPLANT is outlined. Sound is picked up by a lapel microphone and sent for processing to a box worn by the patient and equipped with standard hearing-aid controls for tone and loudness. An automatic-gain-control circuit on a microelectronic chip in the box first reduces the wide fluctuations in loudness that characterize normal ambient sound to the much more limited range suitable for electrical stimulation of the inner ear. A set of band-pass filters then divides the complex alternating-current waveform into four frequency channels corresponding to the major formants produced by the human vocal system in making vowel sounds. The resulting narrow-band signals (one of which is zero in this example) modulate the amplitude of four independent radio-frequency oscillators mounted in the antenna-coil assembly. The transmitting antennas are held in place on the scalp over the implanted receiver coils by mating pairs of ceramic magnets. Each receiver circuit acts as an independent AM-radio set to demodulate the carrier frequency and recover the original narrow-band signal for direct transmission to a pair of stimulation contacts on the electrode, which is inserted in the tympanic canal. Such a scheme is limited to a fairly small number of channels. A group at Stanford University has recently developed an improved eight-channel system in which the detailed characteristics of each channel's output signal are digitally encoded for transmission on a single carrier frequency and decoded in the implanted receiver circuitry.

lip-read. Most patients with either congenital or adult-onset deafness have been enthusiastic users of even the simplest single-channel prostheses.

Efforts are now under way to improve the quality of the sound by more sophisticated preprocessing and better fidelity in the transmission system that serves to convey the signals from the patient's control unit to the implanted electrodes. One important development has been the realization that most, if not all, of the single-channel capabilities can be realized with extracochlear stimulation in the middle ear. This finding bodes well for the conservative but important clinical application of the technology to many patients (particularly children) during a period of uncertainty over evolving and conflicting multichannel designs. It is generally acknowledged, however, that single-channel stimulators cannot deliver enough information to the nervous sytem to enable a patient to converse without visual aids.



SPREAD OF EXCITATION from a given stimulation site within the cochlea to distant parts of the auditory nerve is the most critical factor limiting the number of channels in a multichannel cochlear prosthesis. The problem is particularly troublesome in the case of monopolar contacts (*left*). Increasing the intensity of the stimulation current delivered by such a contact causes the current to spread in a radially symmetric pattern within the fluid-filled chambers of the inner ear, allowing the excitation to reach parts of the auditory nerve that normally serve different places on the cochlea and thereby giving rise to different pitch sensations; stimulation currents from other contacts in these regions would in turn interfere with the pattern shown, making the separate stimulation channels noisy and difficult to resolve. One way to confine the spread of stimulation current is to employ radial bipolar contacts, which tend to produce a much more localized pattern of excitation (*right*).

The effectiveness of the various multichannel prostheses now available is less clear. The devices are inherently complex, and only a few patients who use them have been thoroughly studied. Moreover, the ranks of these patients include individuals in whom the cause of deafness varies widely, as does the condition of the remaining auditory nerve.

When a small, local subset of the auditory-nerve endings along the cochlea is activated with a multichannel device (by either low-intensity monopolar stimulation or bipolar stimulation), the noisy sensation takes on a definite pitch. When several physically separated subsets are activated sequentially, the subject has little difficulty ordering them into a musical scale. It is important to realize that the psychophysical term "pitch" refers to a subjective judgment by the observer, usually of a highly complex acoustic waveform that may carry little or no spectral energy at the actual frequency corresponding to the stated pitch. Patients with a multichannel prosthesis have likened what they hear to the quacking of ducks or the banging of garbage cans. Such sounds can be ranked by pitch, but they are not at all like the pure tones that are perceived in the case of a sinusoidal acoustic waveform.

It is not clear whether these complex auditory sensations are combinable in the same way as pure sine waves are combined in the speech-simulation experiments. The normal auditory nervous system has an extraordinary capacity for extracting underlying information from noisy signals and generalizing across only distantly related spectral patterns. That is why one can understand human speech by a bass and by a soprano, at a whisper and at a shout, and from speakers with wide variations of accent, nasality and inflection. In fact the spectrographic presentation of a given word is so complex and variable that even an experienced analyst cannot identify most words from such visual records.

Two approaches to utilizing the place-pitch dimension of acoustic information are now being investigated clinically. One is related to the frequency-channel, or vocoder, simulation of speech sounds, in that it filters the acoustic signal into a set of bands, each corresponding approximately to the pitch perception generated by one of the available bipolar electrode pairs. For eight such electrodes there would be eight independent time-varying activations of the auditory-nerve subsets, each activation conveying information about the instantaneous, relative intensity of sound energy in its own band. In principle the particular amplitude-modulated waveform used to activate each channel should not matter. For several reasons, however, the actual waveforms transmitted by each band-pass filter in the sound processor are usually applied directly to the electrodes. Special preprocessing steps are added to compensate for the narrow dynamic range and the frequency-dependent sensitivity of the neurons to electric currents.

The number of independent, parallel stimulation channels is expected to have a significant effect on the intelligibility of speech. This effect, however, will be apparent only if each such channel actually does activate a spatially well-localized subset of the auditory-nerve fibers over the dynamic ranges and temporal patterns typically encountered in normal hearing.

At least five separate teams of investigators have studied patients equipped with between three and 12 stimulation channels each. In general, differences in surgical approach, electrode configuration and waveform selection, together with individual patient variability, overwhelm any meaningful comparison. In a few cases a single patient has been tested with systematic variation of the division of the speech spectrum into different numbers of bands and channels. Additional channels lead to dramatic and immediate improvements in word recognition for subjects whose electrodes allow such localized activation. The identification of words from a random sampling has been reported as high as 80 percent with just four-channel stimulation, a level that approaches functional rehabilitation, given the redundancy and contextual cues available in ordinary conversation.

An alternative approach to the

place-pitch encoding employed in most multichannel devices relies on a multichannel electrode array that continuously varies the point of application of a single channel of stimulation. Both the electrodes and the transmission system are considerably simpler in this scheme, but the processing of the speech signal is much more complex. Even closely spaced monopolar electrodes tend to have an orderly sequence of place-pitch sensations when they are activated singly, apparently because the central nervous system has little difficulty identifying the midpoint of a comparatively broad gradient of activated neurons.

The amplitude modulation and frequency modulation of certain parts of the speech spectrum, such as the formants produced by the shape of the mouth cavity, convey particularly useful information for discriminating many acoustic elements of speech. A speech processor based on a microcomputer chip can identify and track one such spectral feature, known as the second formant, and select the best stimulation site and intensity, based on a stored map of the sensations produced by each electrode in a particular subject. The frequency of the stimulation applied at the currently activated site can be used to encode the fundamental frequency of vocal-cord vibration, which corresponds to the low-frequency range for which subjects report a vibrationlike sensation of pitch. A system of this type with 22 individual electrodes, developed by Graeme M. Clark and his colleagues at the University of Melbourne, has been implanted in about 20 patients. Again, the variability among patients overwhelms any attempt to compare results with single-channel electrodes or with multichannel designs, but at least some subjects have achieved significant scores on word-recognition tests.

To improve performance, information channels corresponding to the first and third vowel formants, together with high-frequency consonants, will be needed. Perhaps two or even three simultaneous stimuli might be delivered without significant channel interaction, if they could be kept far enough apart. It remains to be seen whether this general approach to the basic trade-off between the number of channels and their independence will provide the necessary information in a form that is compatible with the information-processing capabilities of the auditory nervous system.

The complexity of the sensations evoked by even the most localized stimulation of the cochlea has surprised and intrigued many investigators. The place-pitch theory in its simplest form predicts that local activation will elicit a fairly pure tonal sensation, corresponding to the local resonant frequency of the basilar membrane. Changes in the frequency and waveform of the electrical stimulus should result only in simple changes in loudness and only to the extent that such changes affect the average rate of firing among the active neurons. How then do spectrally complex sensations such as buzzes and clangs arise? Can they be systematically controlled to provide another form of prosthetic information transfer?

It has long been known that the auditory-nerve activity transmitted to the brain contains rather detailed temporal information about the exact phase of the motion of the basilar membrane. For sound frequencies below 5,000 hertz the exact timing of each neural impulse in an auditory-nerve fiber is locked to the phase of the mechanical motion sensed by the particular hair cell providing its sole input. Even though the fiber must pause for two or three milliseconds between impulses, a spectral analysis of the activity in any single fiber will reveal the frequency of the stimulation that activated it, whether or not the frequency corresponds to the characteristic resonant frequency of that place on the basilar membrane. Some of the central neurons receiving an input from the auditory nerve have a specialized synaptic structure that can transmit and preserve such high-resolution temporal information in spite of the limitation on the overall firing rate of about 300 impulses per second. Furthermore, subjects in psychophysical studies appear to be able to extract precise spectral information even when the loudness of the acoustic stimulus causes the firing rate to be saturated over extended regions of the cochlea. In recent years Murray B. Sachs and Eric D. Young of the Johns Hopkins University School of Medicine have emphasized the potential importance of this temporal information for the discrimination of vowel sounds. Their findings suggest that the nervous system might well be capable of extracting this phase-locked spectral information, and that the failure to correctly reproduce such temporal patterns may give rise to complex, noisy sensations.

What are the critical temporal cues? If many phase-locked neural signals, each with a low overall firing rate, are combined, the frequency of the acoustic source signal can be reconstructed. This approach, however, begs the question of how any particular central neuron receiving such high-frequency driving senses what the frequency is. One way to think of a receiver neuron that is selectively tuned to a particular repetition frequency of its input pattern is to imagine that it is designed to operate by a technique known as temporal autocorrelation. In this approach the input signal is delayed by a time equal to one cycle of the frequency to be detected, and the delayed and undelayed signals are multiplied together. Models of auditory perception often postulate an array of such periodicity detectors in the first few synaptic relays of the brain stem.

Biophysical theory predicts and animal experiments have demonstrated that electrical stimulation with either sinusoidal or pulsed alternating-current waveforms causes significant phase-locking of auditory-nerve fibers for frequencies ranging up to at least 3,000 hertz. Yet prosthesis users consistently report very little change in pitch and only subtle changes in the quality of the sound as the frequency of the stimulus is increased above 300 hertz. Even when the natural resonant frequency of the stimulus site and the repetition frequency of the stimulus waveform coincide, there is no sudden improvement in the weak tonality of the sensation.

Such observations are forcing a reexamination of the class of mechanisms whereby networks of neurons might extract temporal information on a time scale that is much finer than their usual synaptic and conduction processes. One possible model, proposed by Mark W. White, Michael M. Merzenich and me, calls for spatial cross correlation to detect particular configurations of basilar-membrane motion caused by the progress of traveling waves of a given frequency. The process is analogous to the localization of sounds in space based on different signal delays between the two ears; indeed, the two processes might share some of the same neural circuitry. The temporal resolution of both sound localization and phase-locked frequency discrimination is on the order of 10 microseconds.

Obviously whether or not one can prosthetically control these processes and the sensations to which they give rise will depend on the critical input cues for each process. Such control may require finer spatial and temporal control of the electrical stimulation than one can ever hope to achieve. Alternatively, some simple trick of electrode configuration or stimulus timing that has not yet been tried may turn out to do the job.

Damascus Steels

They contained more carbon than most modern steels. Along with skillful forging, this property accounted for the renowned strength and toughness of Damascus swords and for their beautiful markings

by Oleg D. Sherby and Jeffrey Wadsworth

In *The Talisman,* a fictional account of the Christian crusaders' adventures in Palestine, Sir Walter Scott tells of a meeting between Richard the Lionhearted and the Saracen king Saladin. The two adversaries were extolling the virtues of their swords. To demonstrate the strength of his heavy, two-handed straight blade, Richard cleaved a steel mace. In reply Saladin took a silk cushion and "drew [his] scimitar across the cushion... with so little apparent effort that the cushion seemed rather to fall asunder than to be divided by violence." The startled Europeans suspected a trick, but Saladin then emphasized his point by slicing through a limp cloth veil in midair. As Scott describes it, the Saracen's marvelously sharp and wieldy weapon had "a curved and narrow blade which glittered not like the swords of the Franks, but was, on the contrary, of a dull blue color, marked with ten millions of meandering lines."

Although this description contains some poetic license (Saladin's sword could not, for example, have been a scimitar, because curved blades were not introduced until several hundred years after the purported encounter with Richard in 1192), it is a substantially accurate portrait of a type of blade used throughout Islam in Saladin's time. These blades were exceptionally strong in compression—that is, they were hard enough to retain a sharp cutting edge—but they were also tough enough to absorb blows in combat without breaking. They owed their mechanical virtues, as well as their beautiful wavy surface markings, to the material from which they were forged: Damascus steel. By the time of the Crusades, Damascus swords and armor were the stuff of legend. For centuries thereafter they remained objects of fascination and frustration for European smiths, who tried in vain to reproduce consistently the distinctive damask, or surface pattern.

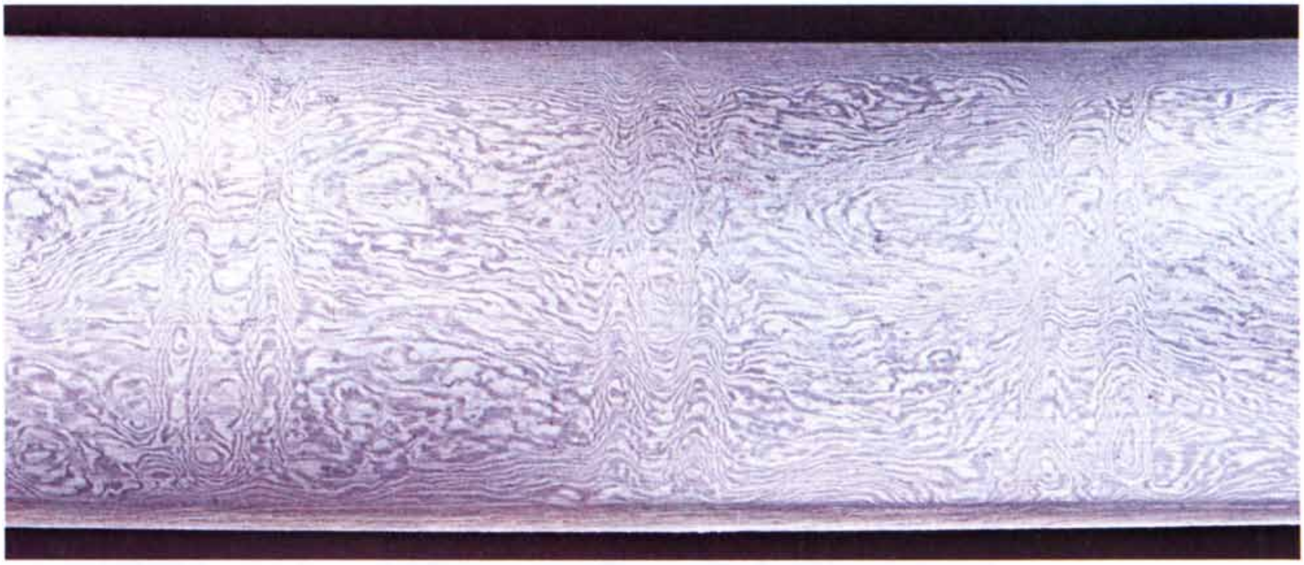
Uncovering the secret of Damascus swords was a challenge that also attracted the attention of eminent European scientists. Among them was Michael Faraday, a blacksmith's son, who attributed the unique properties of a Damascus steel he analyzed in 1819 (before he invented the electric motor and the electric generator) to small amounts of silica and alumina. Although his conclusion was erroneous, Faraday's paper inspired Jean Robert Bréant, Inspector of Assays at the Paris Mint, to undertake a series of experiments in which he added a variety of elements to steel. It was Bréant who first realized, in 1821, the essential metallurgical point about Damascus steels: that their uncommon strength, toughness and beauty arise from their high carbon content. Bréant identified the white areas in the damask as "carbureted steel"; the dark background he referred to simply as "steel."

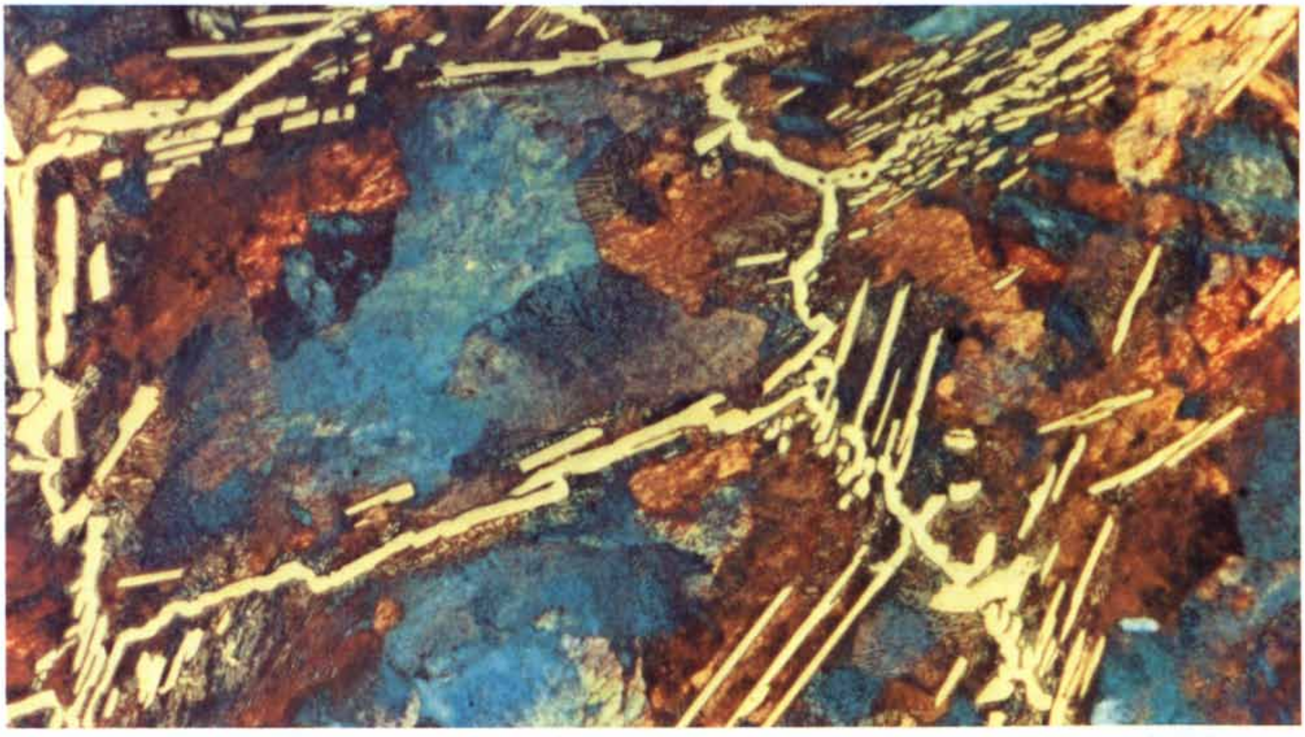
Bréant succeeded in making swords with a damasklike pattern, but he never explained the procedure in detail. Moreover, he could not have understood the importance of all the steps involved. The basis for a complete scientific understanding of Damascus steels was not established until the turn of the century, when a series of investigators worked out the phase transitions that steels undergo as a function of temperature and carbon content. Even today, when the iron-carbon phase diagram is well known, the art of Damascus swordmaking is a patentable discovery under American law.

Our own interest in the subject stems from our research on modern ultrahigh-carbon steels. Such steels, which contain from 1 to 2.1 percent carbon, are rarely used in commercial application because they are assumed to be brittle. The proportion of carbon in Damascus swords, however, ranges from 1.5 to 2 percent. Their undisputed reputation for toughness suggests that the brittleness ordinarily induced by high carbon content can be avoided through proper processing. At Stanford University we have manufactured ultrahigh-carbon steels that, like the swords, are both strong and ductile at room temperature. We have also been able to reproduce the legendary damask. The procedures followed in our laboratory are essentially similar to those invented in the smithies of the ancient Near East.

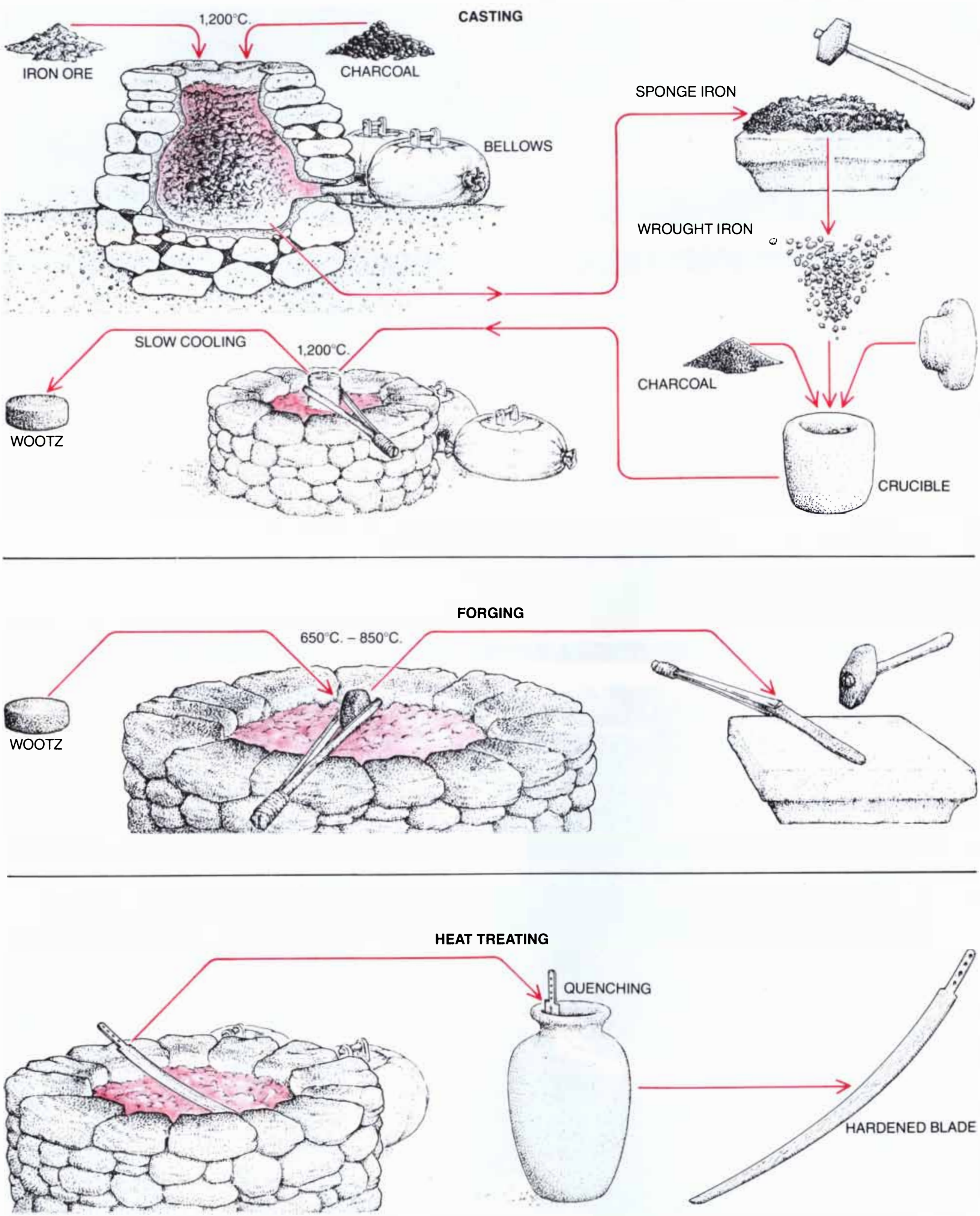
The earliest description of Damascus blades dates from A.D. 540, but they may have been in use as long ago as the time of Alexander the Great (about 323 B.C.). Their name derives not from their place of origin but from the place where Europeans first encountered them during the Crusades. The steel itself was made in India, where it was known as wootz. It was widely traded in the form of castings, or cakes, that were about the size of a hockey puck. The best blades are believed to have been forged in Persia from Indian wootz, which was used to make shields and armor as well. Although the geographic distribution of Damascus steels generally followed the spread of the Islamic faith, they were also known in medieval Russia, where they were called *bulat.*

Like all steelmaking processes, the manufacture of wootz involved the removal of oxygen from iron ore, which is an oxide; the addition of carbon to the reduced iron strengthens it and transforms it into steel. The carbon source was charcoal, wood or leaves. Typically iron ore and charcoal were mixed and heated to about 1,200 degrees Celsius in a stone hearth. Oxygen was removed from the ore by reactions with carbon in the charcoal. Depending on the amount of charcoal in the mixture, the product was a wrought iron, which has a very low carbon content, or a pig iron containing more than 4 percent carbon. The Indian steelmakers manufactured wootz either by





SURFACE MARKINGS on a Persian scimitar (*top and middle*) reflect variations in carbon content within the ultrahigh-carbon steel: the whitish areas of the damask are iron carbide (cementite), and the dark background is iron containing considerably less carbon. The pattern became visible only after the finished blade was polished and etched with an acid that preferentially attacked the iron background. The cementite network is clearly visible in the micrograph of a modern ultrahigh-carbon steel (*bottom*); layers of cementite also alternate with iron layers in the background structure. Damascus steels were toughened by forging, which dispersed the cementite and gave the surface pattern its final form. The effects of forging are recognizable on the Persian sword: the vertical markings in its unusual damask, called Mohammed's Ladder, resulted from local hammer blows. The sword, which dates from the 17th century or later, is in the Metropolitan Museum of Art in New York. The magnification in the micrograph is about 200 diameters.



TYPICAL MANUFACTURING PROCEDURE for a Damascus sword began with the casting of an ultrahigh-carbon steel, called wootz, in Indian foundries. Iron ore and charcoal were mixed and heated to about 1,200 degrees Celsius in a shallow stone hearth. The iron was reduced (stripped of oxygen) by reactions with carbon from the charcoal, and it acquired a spongy consistency. Impurities were expelled from the sponge iron by hammering; the result was bits of wrought iron, which has a low carbon content. The carbon content was then increased by heating pieces of wrought iron with charcoal in a clay crucible, which was sealed to prevent the iron from oxidizing again. When a sloshing sound indicated the presence of some molten matter, the crucible was allowed to cool slowly in the furnace. Wootz was widely traded in the form of cakes several inches in diameter. Near Eastern smiths forged a Damascus blade from an individual cake that was probably heated to between 650 and 850 degrees C.; ultrahigh-carbon steels are ductile in that temperature range. The craftsmen hardened the finished blades by reheating them and then quenching them in water, brine or some other liquid.

adding carbon to wrought iron or by removing carbon from pig iron.

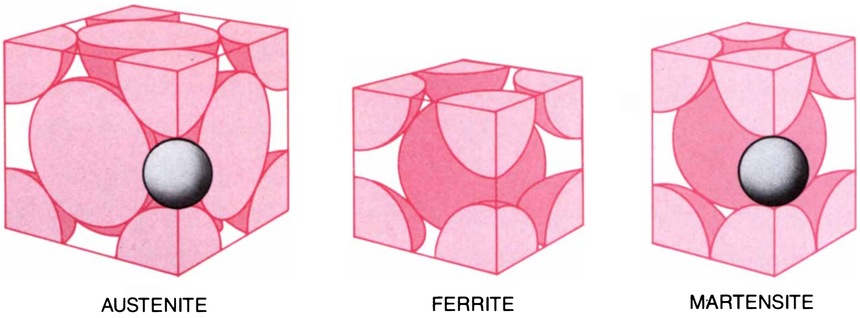
The procedure for producing wootz from wrought iron is better understood than the procedure for producing it from pig iron. Small pieces of the metal were mixed with charcoal in a sealed clay crucible about three inches in diameter and six inches tall. The crucible was then heated to roughly 1,200 degrees C. At this temperature wrought iron is still solid, but its crystals have a face-centered cubic structure that can accommodate carbon atoms in the spaces between the iron atoms [*see top illustration at right*]. Thus carbon diffused gradually into the iron, forming an alloy known today as austenite.

The addition of carbon lowered the melting point of the metal. When the carbon content at the surface of the wrought-iron pieces had increased to about 2 percent, a thin molten layer of white cast iron began to form on each of the pieces. The presence of molten matter was recognized by a "sloshing" sound as the crucible was shaken; it indicated that a significant amount of carbon had dissolved in the iron.

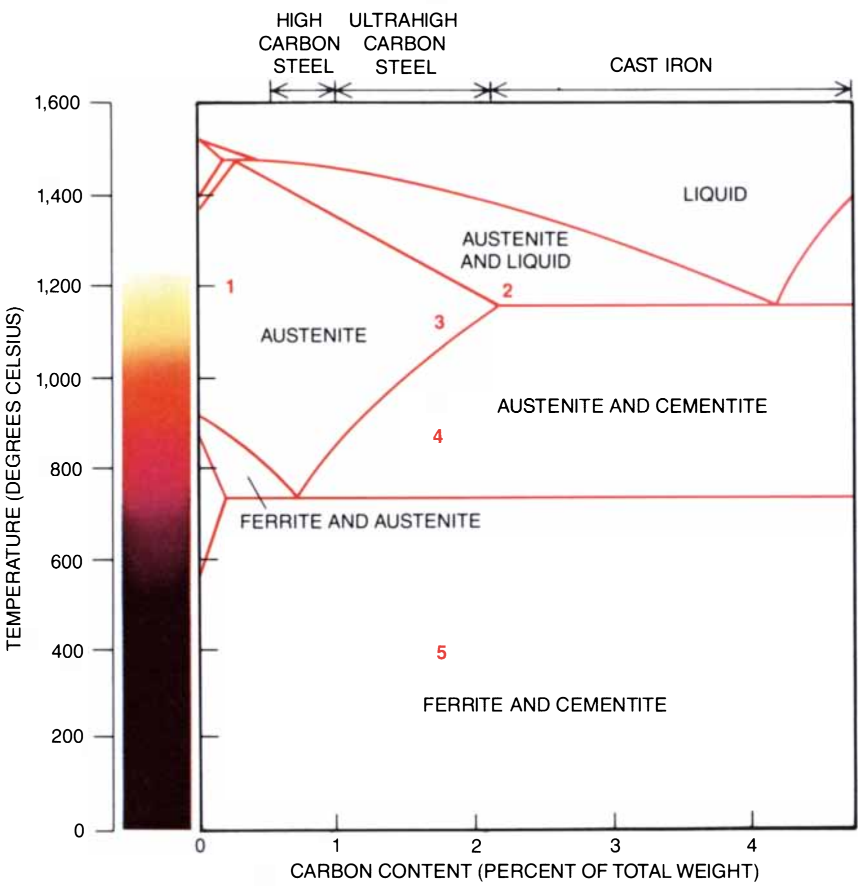
At that point the crucible was cooled very slowly, sometimes over a period of days. The slow cooling produced a homogeneous distribution of from 1.5 to 2 percent carbon throughout the steel. When the temperature of the metal fell below about 1,000 degrees C., some of the carbon precipitated out of solution, forming a network of cementite, or iron carbide (Fe_3C), around the austenite grains. Because the slow cooling allowed the grains to grow relatively large, the cementite network was coarse.

It was this coarse network that ultimately led to the visible markings on Damascus swords. Cementite, however, has certain negative qualities. Although it is very hard, it is also extremely brittle at room temperature. The brittleness would have been made worse by the network structure, which provided ready avenues for crack growth. Yet Damascus swords were the opposite of brittle: they were tough. Wootz acquired toughness only after it had been forged and the cementite network had been broken up by extensive hammering.

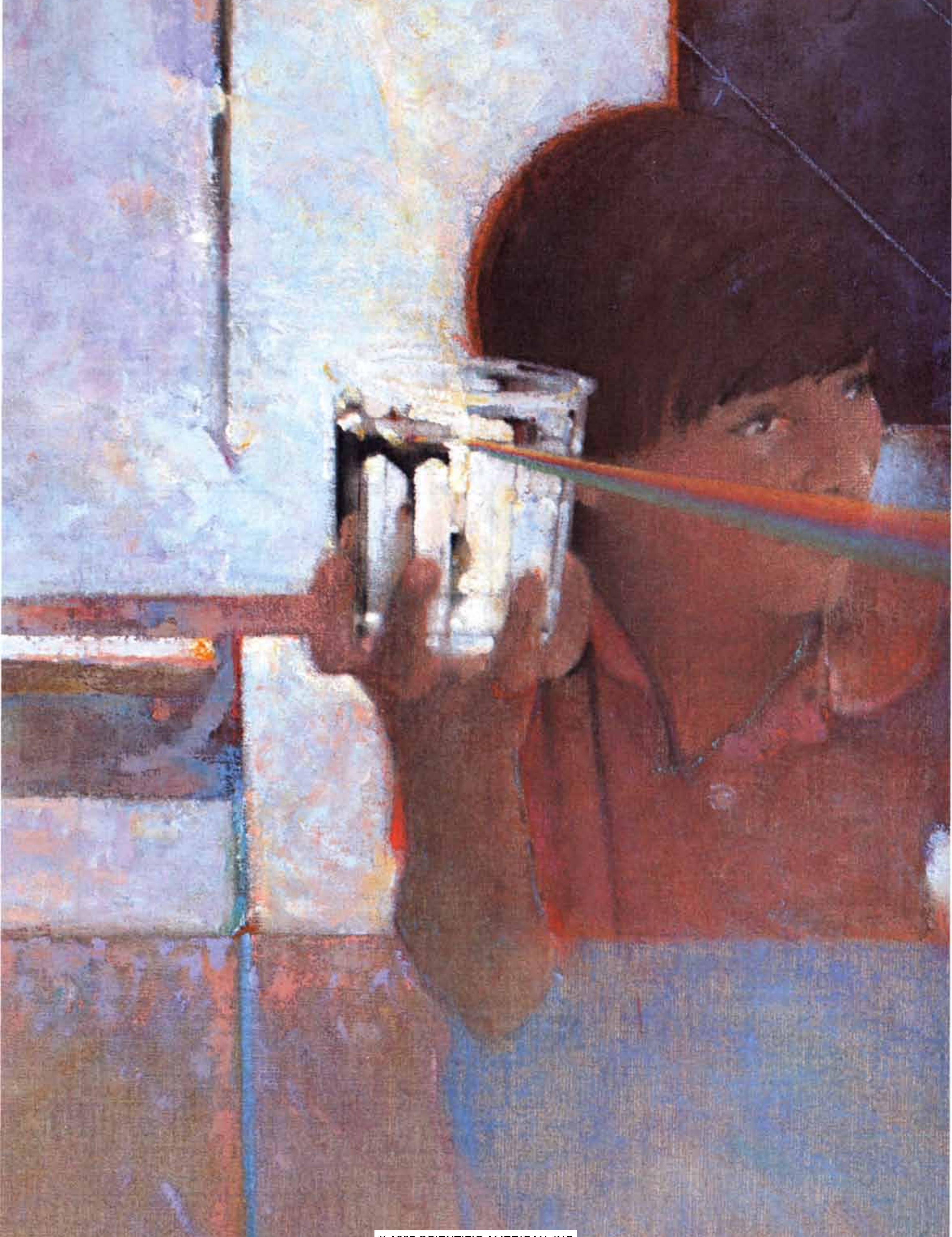
Damascus steels were apparently hammered into blades at a fairly low temperature. Medieval blacksmiths could not precisely measure the temperature inside a hearth or a forge, and so they used the color of the metal as a guide. The forging range for steels typically extended from white heat (1,200 degrees C.) down to orange (900 degrees), but wootz seems to have been

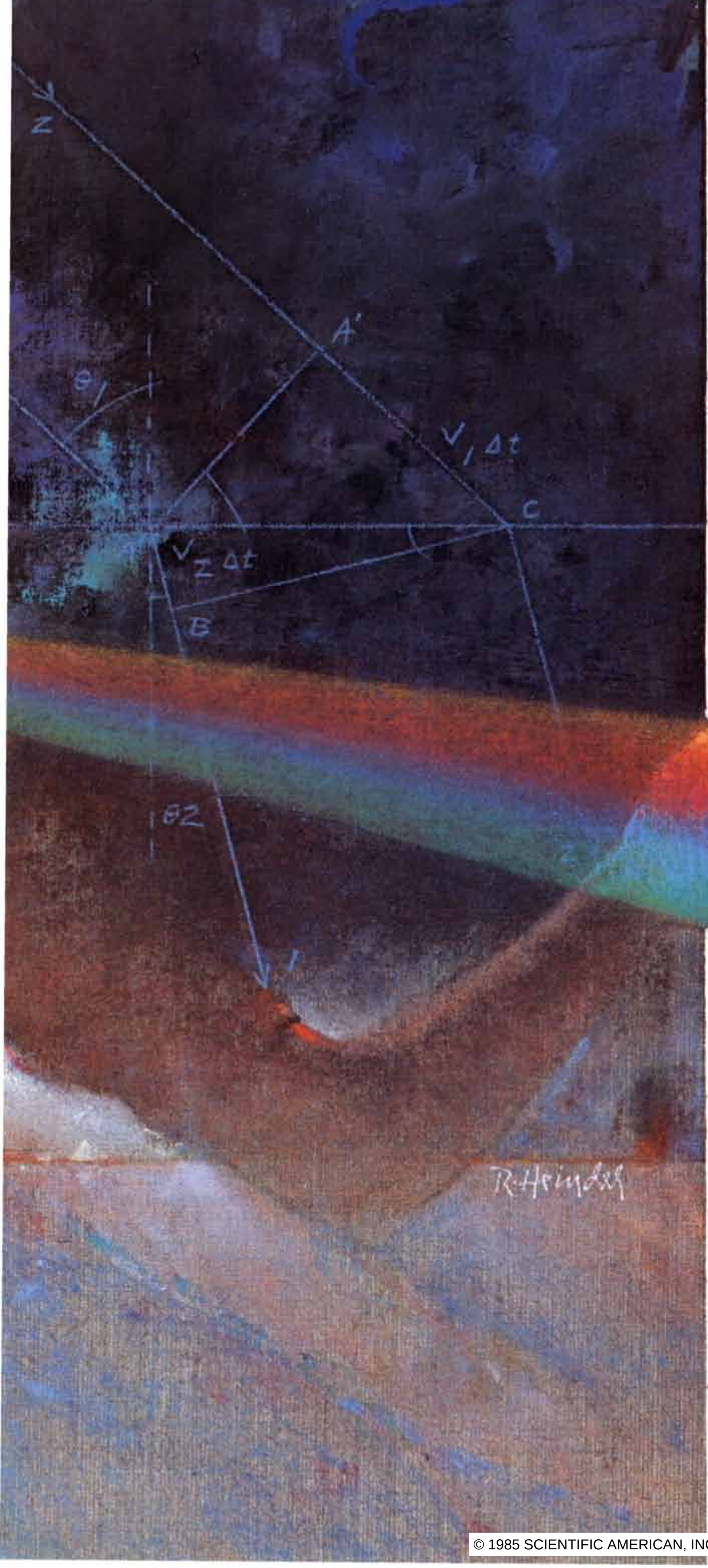


IRON CRYSTALS in ultrahigh-carbon steels can exist in three forms. At temperatures above 727 degrees C. the stable configuration is a cubic lattice with iron atoms at the center of each face. The face-centered arrangement can accommodate carbon atoms (*dark balls*) between the iron atoms. Thus at high temperatures carbon dissolves in the iron; the solution is called austenite. If the steel is then cooled slowly to room temperature, the iron crystals convert into a body-centered cubic form in which there is little room for carbon; this phase is called ferrite. If the steel is cooled rapidly (quenched), carbon atoms are trapped in distorted, tetragonal body-centered crystals, called martensite, that are harder than ferrite.



IRON-CARBON PHASE DIAGRAM is the basis for understanding the properties and the manufacturing process of Damascus swords. When wrought iron and charcoal were heated to 1,200 degrees C. in a crucible, the iron converted into face-centered austenite (*1*). Carbon from the charcoal could then dissolve in the iron, decreasing its melting temperature. Molten cast iron formed at the surface of the iron particles when the carbon content of the surface layer exceeded 2 percent (*2*). Slow cooling allowed the carbon to diffuse through the metal, producing a steel with an average carbon content between 1.5 and 2 percent (*3*); it also allowed the austenite grains to grow to a coarse size. When the temperature fell below about 1,000 degrees C., carbon precipitated out of solution as cementite at the grain boundaries (*4*). The coarse cementite network was the source of the whitish damask markings. As the temperature fell below 727 degrees, face-centered austenite converted into alternating layers of cementite and carbon-poor, body-centered ferrite (*5*). The blades were hardened by being reheated above 727 degrees and then quenched, which converted austenite into martensite. Medieval smiths gauged the metal's temperature from its color.





HOW A MOMENT OF LEARNING CAN INSPIRE A LIFETIME OF ACHIEVEMENT.

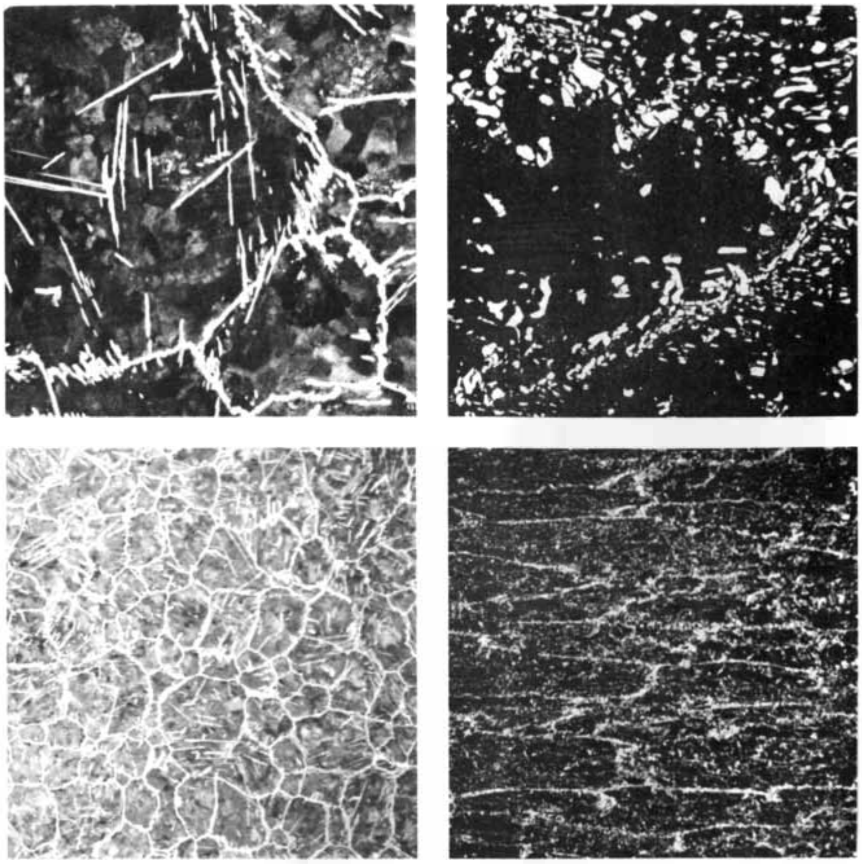
Watch a child the moment a fresh idea begins to take hold. Curiosity gives way to discovery. And with it comes the exhilaration of knowing something special and new.

Thus the reason for a series of films about math, science, and economics, all funded and distributed by Phillips Petroleum. In them, learning is pictured not as a survival skill, but as a lifelong passion that can thrill, inspire, and ultimately, reward.

Since 1976, millions of young people have seen "American Enterprise" and "The Search for Solutions." Shortly, many will be seeing "The Challenge of the Unknown." To find out more about these films, write: Phillips Petroleum Company Public Affairs Department Bartlesville, Oklahoma 74004.

It's our hope that with the help of a little illumination, whether from a film, teacher, or just an experience, young people can be encouraged to think, and do, the impossible.

Phillips Petroleum Company



EFFECTS OF FORGING on Damascus steels were simulated by rolling a modern ultrahigh-carbon steel. The steel is shown magnified 130 times (*top*) and 6.5 times (*bottom*). Before rolling (*left*), the light cementite network is continuous; its dimensions are roughly the same in all directions. After rolling (*right*), the network is elongated in the rolling direction and is broken up into spheroidal particles. As a result the steel is tougher (less brittle).

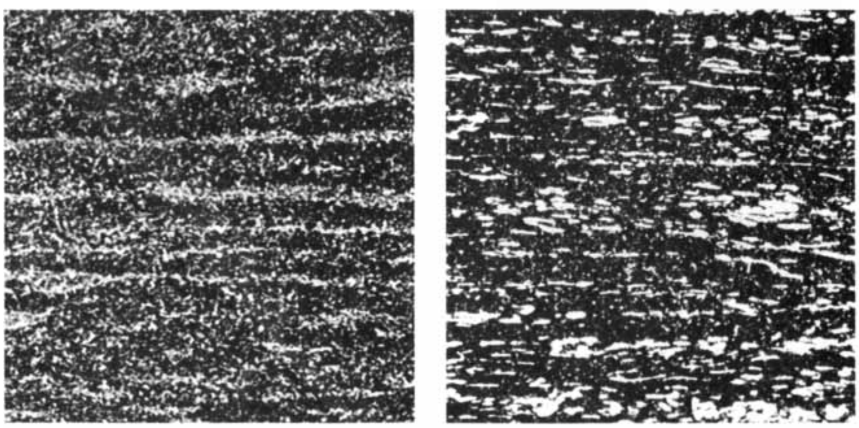
forged at temperatures between cherry (850 degrees) and blood red (650 degrees). Higher temperatures would have caused the cementite to dissolve again in the austenite. Hammering the wootz cake at a temperature below 850 degrees, on the other hand, broke up the continuous cementite network into spheroidal particles. The carbide particles still served the function of strengthening the steel, but because they no longer formed a continuous network, the metal was not brittle.

Damascus swords show evidence of having been well forged; the height of the ingot seems to have been reduced by a factor of three to eight in the pounding out of the blade. An experiment we did demonstrated that ultrahigh-carbon steels are indeed ductile and easily forged at a temperature of 850 degrees C. We subjected ingots with carbon contents of 1.3, 1.6 and 1.9 percent to large strains, compressing them in a single step by a factor of three. None of the ingots showed any signs of cracking. In contrast, a cast-iron ingot, whose higher carbon content (2.3 percent) rendered it brittler, cracked around the edges under the same strain.

One of the reasons European smiths had so much trouble reproducing Damascus blades, even from imported wootz, may have been that they were accustomed to working with low-carbon steels, which have a higher melting point. As a result they may have tried to forge the Indian steel at a white heat, when it would have been partially molten. The likely outcome of such an effort was described by Bréant, who observed that "at a white heat [Damascus steels] crumble under the hammer."

After forging, Damascus blades were usually hardened by heat treatment. Thermal hardening of a steel is achieved by heating it above 727 degrees C. (the temperature at which body-centered ferrite begins to be converted into face-centered austenite) and then quenching it (cooling it rapidly) in water or some other medium. When ultrahigh-carbon steels are allowed to cool slowly from the austenite phase, as in the initial casting of wootz, the austenite is converted into pearlite: alternating layers of soft, carbon-poor ferrite and carbon-rich cementite. If the steel is quenched, however, the transformation of austenite into pearlite is suppressed. The iron crystals become body-centered, but they are stretched from a cubic form into a tetragonal one. This structure, called martensite, still has room for carbon atoms, and so it is hard.

Medieval smiths seem to have followed a variety of procedures in heat-treating Damascus blades, often giving considerable weight to considerations that seem immaterial to the modern engineer. Some smiths, for instance, insisted that swords be quenched in the urine of a redheaded boy or in that of a "three-year-old goat fed only ferns for three days." One of the most detailed descriptions of a procedure for hardening Damascus steel (*bulat*) was found in the Balgala Temple in Asia Minor: "The bulat must be heated until it does not shine, just like the sun ris-



THICKNESS SECTIONS of a Damascus blade (*left*) and of a rolled ultrahigh-carbon steel (*right*) show similar microstructures, suggesting they were processed in a similar way. Mechanical working compresses the cementite network; the space between the layers is about 100 micrometers. Rolled steels have a less intricate damask than the forged blades.

When the weather gives you reason to worry, Eagle gives you good reason not to.

No matter how good a driver you are, a clap of thunder, a darkening sky or a threatening weather forecast can be quite unsettling...unless, of course, you're driving an American Eagle.

With the simple flick of a switch, Eagle gives you the confidence of 4-wheel drive. You can even shift-on-the-fly, a feature you'll really appreciate when faced with a sudden shower, blizzard or stretch of bad road. No wonder the National Ski Patrol has chosen Eagle as its Official Vehicle.

NATIONAL SKI PATROL

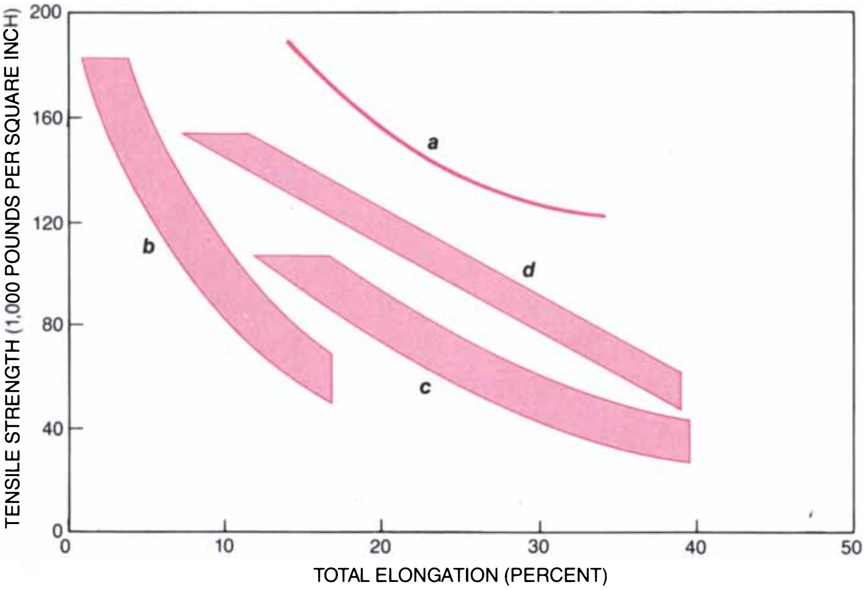
Of course, confidence isn't the only luxury you'll enjoy in an Eagle. With features like power steering, power windows and door locks, plush carpeting and reclining seats of genuine leather, Eagle seats five in total luxury.

The American Eagle. It's the luxury wagon that takes much of the worry out of bad weather driving.

The 2-wheel/4-wheel drive American Eagle

AVAILABLE FOR PURCHASE OR LEASE AT PARTICIPATING AMC DEALERS.

SAFETY BELTS SAVE LIVES.



STRENGTH AND DUCTILITY of ultrahigh-carbon steels can be greater than those of conventional steels. Tensile strength is the maximum tensile stress a standard piece of metal can withstand before fracturing. Elongation is a measure of ductility: it is the amount by which the metal can be stretched before fracturing. When ultrahigh-carbon steels (*a*) are rolled during cooling to produce an ultrafine microstructure, they are stronger (for a given ductility) than conventional low-carbon steels (*b*) and high-strength steels containing small quantities of special alloying elements (*c*). They also outperform certain advanced steels (*d*).

ing in the desert, after which it must be cooled down to the color of king's purple, then dropped into the body of a muscular slave... the strength of the slave was transferred to the blade and is the one that gives the metal its strength."

The instructions may be reinterpreted as follows. The blade was heated to a high temperature, presumably above 1,000 degrees C. ("sun rising in the desert"), and then air-cooled to a temperature of about 800 degrees ("king's purple"). Finally, it was quenched in a warm (37 degrees), brinelike medium.

This particular procedure would probably not have yielded the best Damascus blade. Heating the blade above 1,000 degrees C. would allow the cementite to dissolve again in the face-centered austenite; on being cooled to 800 degrees, the coarse network structure, which had been eliminated by forging, would reemerge. The high temperature would also have generated relatively large steel grains. Both effects would have reduced the toughness of the blade. A sword manufactured according to the Balgala prescription would have been hard, but it might also have been too brittle to survive the impact of a Damascus blade that had been heated to only just above 727 degrees before being quenched. Such a blade would have been tough as well as hard.

In general, according to modern metallurgical theory, the strongest and toughest steels are those with the finest grains and particles. Ironically, this suggests that the best Damascus swords of all might have been ones that lacked a damask pattern. Medieval smiths undoubtedly used the characteristic surface markings as a form of quality control: the damask was evidence both of a high carbon content, which made the blade strong, and of a well-forged structure, which made it tough. A damask would only be visible, however, if the cementite particles in the steel were coarse and unevenly distributed. Blades whose microstructure was so fine that it produced no visible surface markings would have been stronger and tougher still.

In order to test our ideas about the composition and processing of Damascus steels, we attempted to reproduce the damask in the laboratory. First we heated a small steel casting, whose carbon content was 1.7 percent, to a temperature of 1,150 degrees C. (light yellow) for 15 hours. The prolonged heating dissolved the carbon and produced a very coarse austenite. Next the casting was cooled at a rate of about 10 degrees per hour. The slow cooling allowed a coarse, continuous cementite network to form at the austenite grain boundaries.

Finally we reheated the casting to 800 degrees C. and rolled it, reducing its height by a factor of eight. This step, which simulated forging, stretched the grains in the rolling direction and broke up the carbide network. When the steel was etched with an acid that attacked the iron matrix preferentially over the carbide, a damask was visible to the unaided eye. The microstructure of the casting was strikingly similar to that of Damascus steels [*see bottom illustration on page 118*].

The procedure described here is one way of manufacturing a Damascus steel; there probably were many others. The Near Eastern artisans may even have made superior ultrahigh-carbon steels that lacked a damask. We did just that in the laboratory by rolling the casting as it cooled from 1,100 degrees C. through the austenite-plus-cementite phase. The mechanical working refined the austenite grains and caused the cementite to precipitate out of solution as fine, uniformly distributed particles rather than as a coarse network. Thus the finished steel bore no surface markings.

Such damask-free ultrahigh-carbon steels are stronger and more ductile at room temperature than conventional automotive steels. Furthermore, they are superplastic (that is, they behave like molasses or semimolten glass) at temperatures of 600 to 800 degrees C. As a result they can be shaped with precision into complicated objects, such as gears, with a minimum of expensive machining and by processing methods that are adaptable to mass production. This suggests they might find wide industrial application.

We are not the first to claim to have rediscovered the lost art of Damascus steelmaking. In addition to Bréant and Faraday, one of our predecessors was the Russian engineer Pavel P. Anosoff, who in 1841 published a two-volume monograph titled *On the Bulat*. So enthusiastic was Anosoff about his findings that he proclaimed: "Our warriors will soon be armed with bulat blades, our laborers will till the soil with bulat plowshares, our artisans will use tools fashioned of bulat, and bulat will supersede all steel employed for the manufacture of articles of special sharpness and endurance."

His forecast was not realized. Today the enormous potential of ultrahigh-carbon steels remains largely unexploited. Although our optimism is more tempered than Anosoff's, we believe the situation will change and the secret of Damascus steel will become common knowledge in modern industry. In the words of an old Russian proverb, "The best of the new is often the long-forgotten past."

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $245 for the first 12 months (current renewal is $185); otherwise, return the books; there is no charge.

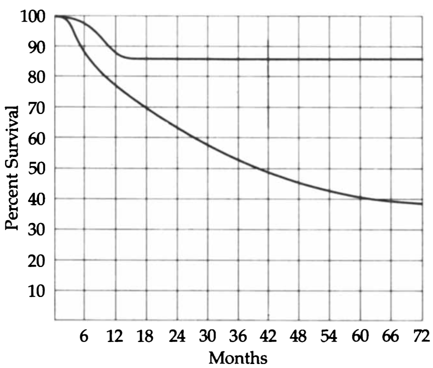
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

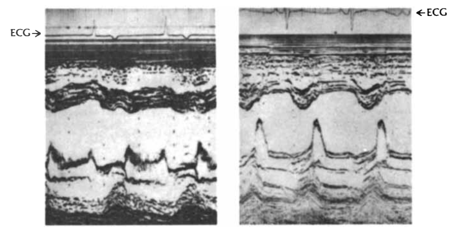
*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

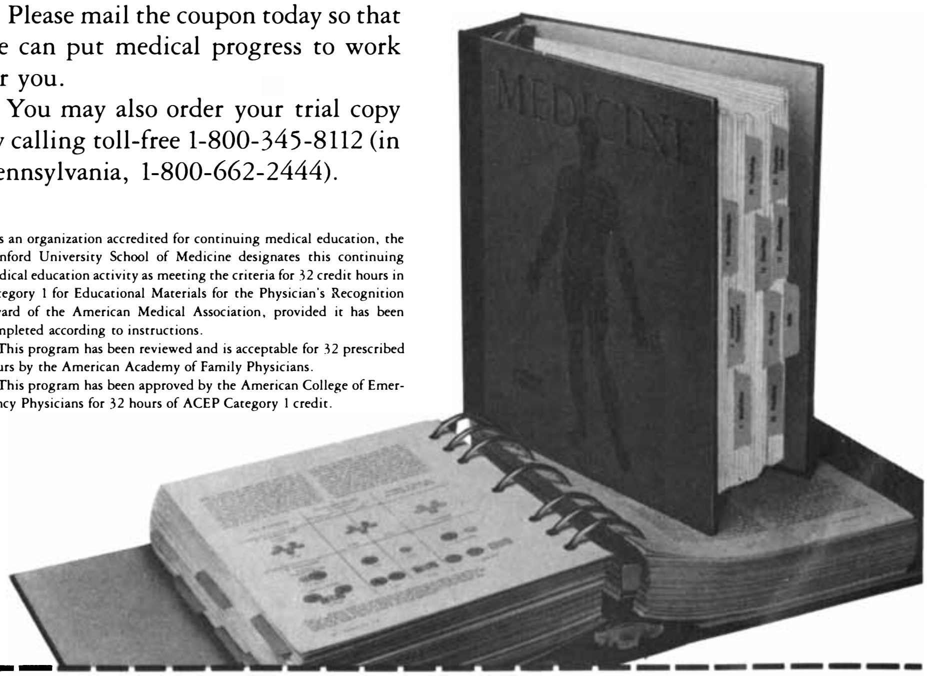
This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

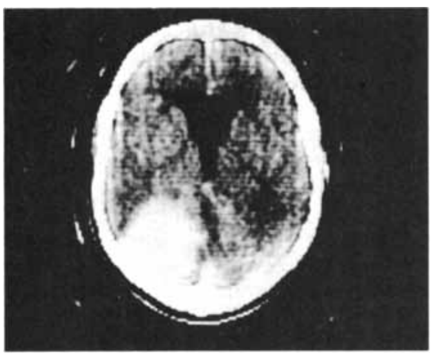


Top curve represents survival in chronic hepatitis patients treated with steroids.



Echocardiograms from patients with aortic regurgitation.





CT scan reveals large meningioma and mild hydrocephalus.

SCIENTIFIC AMERICAN MEDICINE

415 MADISON AVENUE, NEW YORK, NEW YORK 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $245 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $245*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date________ Account Number________

*Please add sales tax for California, Illinois, Michigan, Ohio and New York

Name________ MD Specialty________

Address________

City________ State________ Zip________

Signature________

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

7T

THE AMATEUR SCIENTIST

Bidwell's ghost and other phenomena associated with the positive afterimage

by Jearl Walker

Adapt your eyes to indoor darkness for 10 or 15 minutes and then set off a bright flash of light. Keep your gaze steady as the glare subsides. Within a few seconds you will see what resembles a snapshot of the room. Details lost in the dazzle of the flash are now readily apparent. This is an afterimage (a positive one because it reproduces the light and dark regions in the scene that was illuminated). It lasts for tens of seconds if you keep your eyes steady.

I first learned of this kind of afterimage from Peter Hasselbacher of the University of Louisville School of Medicine, who had seen it at parties when he was in college. One of the earliest reports of the afterimage was made in 1894 by Shelford Bidwell, a British physicist who did research on vision. When he illuminated an object with the flash from an electric discharge, he perceived the object in an afterimage about .2 second later. Sometimes the afterimage faded and reappeared several times.

In another experiment Bidwell made a gas-discharge tube rotate at about .4 revolution per second. The tube was periodically illuminated with a flash of light. When he viewed the rotation with a steady gaze instead of tracking it with his eyes, the tube appeared to be followed by a dim blue or violet ghost light. When the rotation stopped, the ghost light caught up with the discharge tube. The afterimages Bidwell saw in each of his experiments have since been called Bidwell's ghost.

Bidwell also shone a light on the hole of a rotating disk so that a spot of light circled a screen on the other side of the disk. The room was otherwise dim. He perceived a ghost light following the spot at a delay of about .2 second. When the original light was white, blue or green, the ghost light was violet. When the light was orange or yellow, the ghost light was blue or blue-green. Red light, however, did not produce a ghost.

In 1934 William Stewart Duke-Elder, a British surgeon and ophthalmologist, described how an afterimage can be produced while an observer looks toward a spot of light with a card blocking the view. When the card is rapidly moved in and out of the line of vision, the observer perceives a positive afterimage of the light. "So vivid, indeed, may be the impression of the original afterimage," Duke-Elder wrote, "that the card appears transparent, and details which were not noted in looking at the light are brought to the attention in the afterimage."

I investigated the afterimage with an electronic flash from my camera. After a few minutes in a dark room (well before my eyes were adapted to the dark) I opened a magazine on a table. The room was so dark that I could see nothing of the magazine. Having placed the flash to one side of my head so that it pointed toward the magazine, I triggered it. (I was careful not to point the flash unit directly into my eyes. At such close quarters the light might damage the retina.) The glare was dazzling, but I was able to make out the general outline of the magazine pages.

Within seconds a positive afterimage of the magazine appeared. (I call it a snapshot afterimage.) The image was bright enough for me to recognize photographs and read the headlines. I "saw" the smaller print but could not read it. The image soon blurred and faded, finally disappearing after about 15 seconds. The darker sections (the photographs) disappeared first and the lighter sections last.

What came next was a vaguer but much more persistent negative afterimage in which the bright and dark regions were reversed. I believe the negative afterimage results from a retinal fatigue caused by the snapshot afterimage. The bright portions of the illuminated scene create much activity in the regions of the retina where their images fall. When that activity dies out, those regions seem dark compared with the areas that had less activity. Hence the positive afterimage fades into a negative one.

I continued flashing the light and examining the snapshot afterimages for the next half hour. For the first 15 minutes they improved steadily in quality as my eyes continued to adapt to the darkness. Both the snapshot and the negative afterimages were always colorless even when the objects in view were colored. Once the snapshot afterimage was so vivid that I tried walking across the room as if I could really see. I quit when I walked into the edge of an open door that seemed to be distant.

Improving the quality and duration of the snapshot afterimage took some practice because the flash tended to make me blink. A blink or a movement of the eyes across the scene usually weakened or erased the snapshot afterimage. Moving my head did not seem to matter if I did not appreciably move my eyes with respect to my head.

I checked the steadiness of my gaze during the afterimage by looking at a dark television screen. The flash of light made the screen glow faintly. I then perceived the glowing screen superposed on the snapshot afterimage of the dark set. The two images slowly separated, indicating that in the darkness after the flash I had unwittingly let my eyes drift, apparently without affecting the afterimage, which was fixed in place on my retina.

Nearly all the literature on this type of Bidwell's ghost says the gaze must be kept steady if the afterimage is to be maintained. With a few trials I discovered that a steady gaze is usually not crucial in the early stages of the proceeding. If I move my eyes during the flash of light or in the next several seconds, the afterimage disappears momentarily but soon returns, apparently as sharp as it would have been if I had kept my eyes still. If I move my eyes later, the afterimage disappears immediately and permanently. No one has been able to explain how such a movement of the eyes erases the afterimage.

The negative afterimage was much less fragile, appearing even if I could not keep my gaze fixed. It usually persisted for several minutes. If I flashed the light in order to produce another snapshot afterimage, an old negative afterimage would sometimes be superposed on it.

The perception of a snapshot afterimage often conflicted with my common sense. Once I illuminated my hand while it was in front of my face and then moved it behind my back. When the snapshot afterimage appeared, I had a clear perception that my hand was in front of me and an equally clear feeling that it was behind my back.

In another experiment I squatted, looking toward my feet and the floor while I triggered the light. Then I stood up, while continuing to look at the floor. The snapshot afterimage convinced me that my head was near my feet, but I knew that I was standing up.

The reverse arrangement was equally unsettling. I flashed the light toward the floor while I looked down from an upright position. Before the afterimage appeared I squatted. The contrast between what I perceived and what I felt was startling. Just as disconcerting was a snapshot afterimage of my face when I looked into a mirror during the flash of light.

By triggering the flash unit twice within a few seconds I could superpose two snapshot afterimages. In one experiment I held my hand in front of my face while I flashed the light. Immediately I moved my hand to the right and triggered the light again. When the snapshot afterimages appeared, I saw both positions of the hand. I knew the hands were mine, but neither image made sense because my hand was then at my side. If I held my hand with the palm toward me in one flash and away from me in the next, I saw a remarkable hand with 10 fingers, half curled toward me and half curled away.

I next sat with my right leg crossed over my left. With my head down I illuminated my legs with a flash, quickly crossed them the other way and illuminated them again. In the resulting afterimage I perceived a leg extending to the left and another one extending to the right.

The ability to superpose snapshot afterimages enabled me to stroboscopically freeze an object moving through my field of view. I held a coin in front of me, released it and then flashed the light twice while the coin was falling. In the afterimage I saw my hand and two positions of the coin.

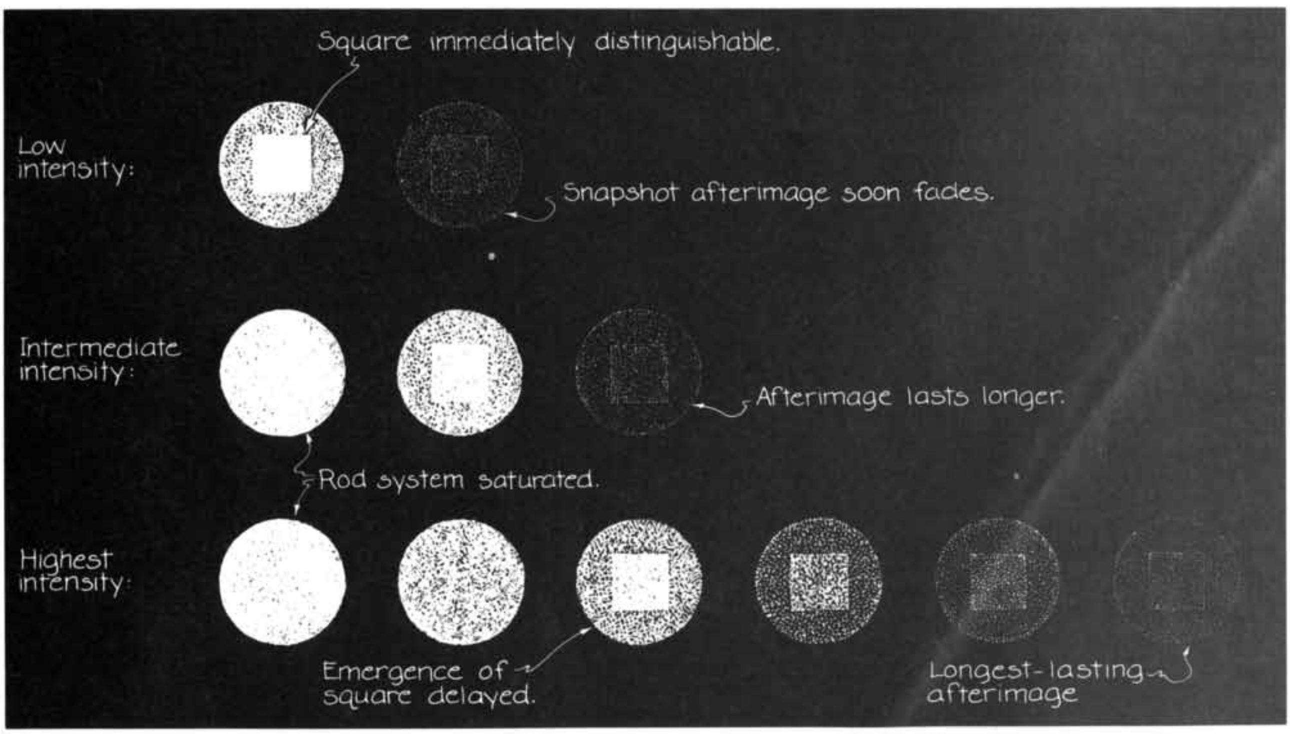
I wondered if an afterimage from one eye could be superposed on an afterimage from the other. I triggered the light while only my left eye was open. Then I closed that eye, opened the other one and triggered the light again. The two views were indeed superposed.

If I moved my head between one flash and the next, the scene was usually too cluttered to be clearly recognizable. Sometimes I created two images that I could fuse to produce a weak stereoscopic effect. Normal stereopsis yields a perception of depth that is correct because the brain is accustomed to the usual separation of the eyes. In my experiment the views from the eyes were made with a different separation, giving rise to a different perception of depth.

I next experimented with several cellophane filters to check for color dependence in the snapshot afterimages. I covered the flash unit with a filter, but I could just as well have looked through the filter. With a green or blue filter the snapshot afterimages were vivid. With a red filter they were dim and brief or did not appear at all.

I then investigated the ghost image reported by Bidwell for a moving spot of light. I moved a small orange spot of light (the trigger button of my camera flash) rapidly across my field of view while I kept my gaze fixed. After a noticeable delay a dim blue spot raced along the same path, running into the orange spot where I had stopped it. The trail behind the ghost spot glowed dimly (it may have been blue, but I was not sure) for tens of seconds. When I moved the orange light through an intricate pattern, the delayed ghost spot raced along the same path, leaving the entire path glowing slightly.

If the ghost image were due merely to the persistence of vision, it would appear simultaneously with the stimulus light, and the path just behind the stimulus would glow. Hence the ghost is an afterimage. As the spot of light is brought across the field of view, its image on the retina moves across photoreceptors that were previously unilluminated. The sudden supply of light



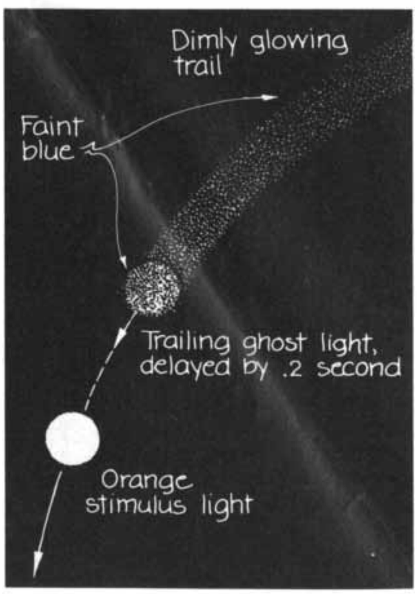
Variations of the positive afterimage

provokes the afterimage, but only after a short delay. If you track the stimulus light, the same photoreceptors are continuously illuminated and no trailing ghost appears.

Returning to my camera flash, I illuminated a page of print with only my right eye open. I was careful to keep gazing toward the left side of the page. Looking at the snapshot afterimage with both eyes, I noticed that the image was poorer on the left side. In normal viewing an object looked at produces an image on the fovea, a small area of the retina that is densely packed with cones. The region lacks rods, the other type of photoreceptor. Because the snapshot afterimage of the page was unclear in the area on which my eye focused, I concluded that rods must produce the afterimage.

There should be a second unclear spot in an afterimage. Toward the temple side of the field of view the eye has a blind spot because that is the part of the retina where the nerves enter the eye. The area has no photoreceptors and cannot transmit an image to the brain. In normal viewing the blind spot is ignored because the region not seen by one eye is seen by the other. When I exposed only one eye to the briefly illuminated page, I expected the afterimage to have an indistinct area toward my temple. In some cases I thought I perceived such an area, but I was never quite certain.

Sometimes I replaced the flash unit with a standard stroboscope set to a low frequency. After a flash or two I turned off the stroboscope to examine the afterimage. (Do not look directly into a flashing stroboscope. The light might be intense enough to harm your eyes. If you are subject to seizures when viewing flashing lights, avoid a stroboscope.)



Bidwell's ghost

In one experiment I generated a snapshot afterimage with my camera flash unit, closed my eyes and then turned on the stroboscope for about 30 seconds. Light came through my eyelids with each flash, creating a diffuse red background (red because of the blood in the eyelids). Against the background I perceived a negative image of the scene that had been illuminated by the camera flash unit. When I turned off the stroboscope, I was surprised to find a snapshot afterimage long after it should have disappeared.

With more experimentation I confirmed that this technique prolonged the snapshot afterimage. I do not know why, but I can offer two hypotheses. The diffuse light transmitted through the eyelids might boost the chemical changes induced by the original flash, making them persist longer. Alternatively, what I perceived after turning off the stroboscope might have been an afterimage that was the contrast of the negative one that would have been there anyway. Since the contrast to a negative image is positive, I perceived a snapshot afterimage when the stroboscope was turned off.

In some classic experiments done with other types of afterimages an observer can vary his judgment of the size of the afterimage if he sees it against a background that changes distance. Suppose he perceives something occupying one degree of arc in his field of view. His decision about the size of the object is determined in part from the apparent distance to the object. If enough clues convince him that the object is distant, he concludes it must be large. If the clues convince him that the object is close, he concludes it must be small.

I had a vague sensation of how the size of the snapshot afterimage can depend on such clues. In a dimly lighted room I illuminated a food can with my camera flash. Then I held a sheet of paper in front of my open eyes. The snapshot afterimage of the can was superposed on the dim outline of the page. When I moved the page away from me, the can did seem to be slightly larger. The experiment was difficult because I tended to focus my eyes on the page. The eye movement erased the afterimage.

Other research on the snapshot afterimage suggests that such a variation in the judged size of an object can arise in total darkness. Supposedly the observer alters the judged size as he moves his hand (unseen in the darkness) from just in front of his eyes to full arm length. I suppose one imagines the hand to be a viewing plane for the afterimage even though the position of the hand can only be felt. I failed to create this illusion.

The positive afterimage following a brief flash of light has been recently investigated by Edward H. Adelson of RCA's David Sarnoff Research Center, drawing on work done independently by Barbara Sakitt of the Massachusetts Institute of Technology and Wilson S. Geisler III of the University of Texas at Austin. Adelson attributes the afterimage to the retinal rods. In his experiments an observer adapted to darkness was exposed to a green square on a circular red background. The flash of light lasted for .01 second. The square occupied 4.5 degrees in the observer's field of view; the background occupied 11 degrees. The patterns were centered about 15 degrees off the observer's line of sight, and so their images on the retina were away from the fovea.

The afterimage depended on the intensity of the light. Some of the results are depicted in the illustration on the preceding page. The top row is a sequence of the observer's impressions when the intensity was low. The square was immediately visible as the glare of the flash subsided, and the afterimage was brief. The second row, a sequence resulting from intermediate intensity, indicates a brief delay before the square could be distinguished. This time the afterimage lasted longer.

The sequence in the bottom row is from high-intensity light. Several seconds went by before the square emerged from the background. The afterimage lasted longer than the others. It is evident that the observer's ability to distinguish parts of the afterimage is delayed when the intensity of the flash is increased. The increase also gives rise to a longer afterimage.

Adelson found in addition that both patterns may seem to brighten as the square emerges from the background. I saw all these results in my experiments with snapshot afterimages of magazine covers and pages.

Adelson suggested how the rod system can be saturated by the flash of light. When the outer segment of a rod absorbs light, a series of chemical reactions generate a substance blocking the flow of sodium ions through a membrane in the outer segment. When the sodium flow is totally blocked, the signal sent to the brain is at maximum strength and the rod is said to be saturated. During this stage additional light will not augment the signal.

Hence the initial flash of light can

saturate the rod system so that the signal from the rods stimulated by the square is as strong as the signal from the rods stimulated by the background. As the signals decay, however, they eventually differ in strength, enabling the observer to distinguish the square. When the flash of light is intense, the rod systems remain saturated longer, delaying the emergence of the square.

In one series of experiments Adelson tested the color dependence of the positive afterimage. The background was kept red-orange while various colors were chosen for the square. With each trial the observer adjusted the intensity of the square so that in the afterimage the square was barely distinguishable from the background. When the choices of intensity were plotted against the wavelength for the color of the square, the graph matched the wavelength dependence of rods, indicating that the rods were responsible for the afterimage.

Adelson also ran trials in which the color of the square was kept green while the background color was varied. This time the observer adjusted the intensity of the background until the square was just visible in the afterimage. Again the graph of the results matched the wavelength dependence of the rods.

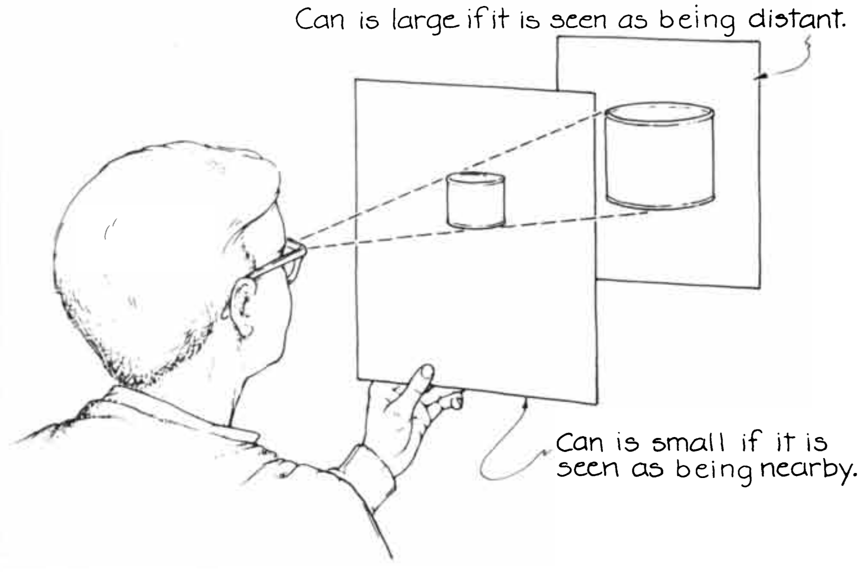
Are rods also responsible for the ghost light trailing a moving stimulus light? I think they are, but I am puzzled by the color of the trailing ghost.

I can suggest a possible solution. The color of the ghost may come about because rods interfere with the color signal from cones during the afterimage. (Rods cannot directly signal color but cones can.) Bidwell's observation that a red stimulus light does not generate a ghost is probably attributable to the fact that rods are not excited by deep red. Therefore such a stimulus does not create the chemical activity in rods that generates a ghost.

A stimulus light of any other color excites the rods so that they generate an afterimage. Their activity inhibits the color signal (then weak but still present) from the cones that were excited by the stimulus light. Such an inhibition sends a signal to the brain that the cones are intercepting the complementary color of the stimulus light.

Suppose the stimulus color is yellow. It will excite the rods and those cones designed to detect yellow. As the stimulus light passes over a region of these cones yellow is perceived.

Within about .2 second the rods create an afterimage. Their activity inhibits the signal of yellow from the cones; a signal of blue (the complementary



Variations in the apparent size of an object in an afterimage

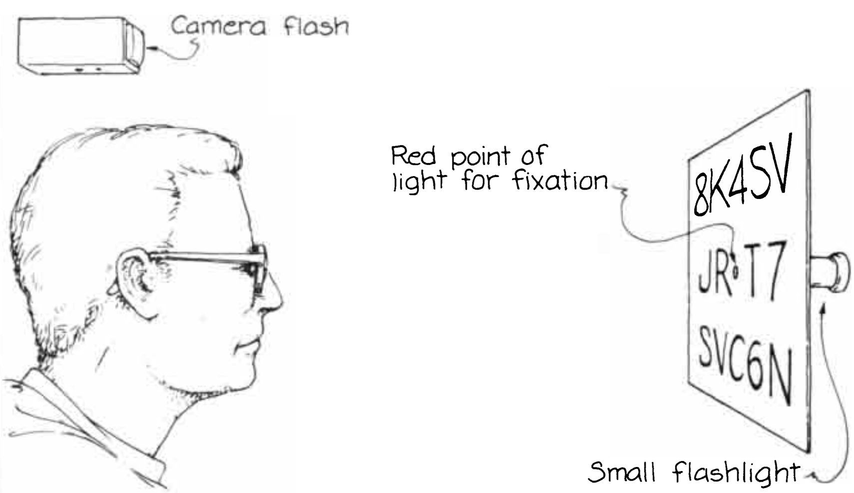
color of yellow) is sent to the brain. The trailing ghost created by the rods is colored blue by their interference with the cones. That subtle coloring is lost if the rods are saturated, as they are when a snapshot afterimage is made. I checked this conclusion by searching for a trailing ghost from an orange stimulus light just after I created a snapshot afterimage. I saw the orange light but no ghost.

Daniel H. Kriegman and Irving Biederman of the State University of New York at Buffalo have recently investigated the extent of information available in the snapshot afterimage. Earlier investigators had tested an observer's recall after glimpsing a display of 12 letters under normal lighting conditions. Usually an observer can recall only three or four letters. The rest are lost because the visual image rapidly fades and because the memory is inadequate.

Kriegman and Biederman wondered if recall would improve when the observer was adapted to darkness and was then shown an array illuminated with a flash of light triggered by the observer. The brightness was adjusted by filters of neutral density placed in front of the flash unit.

The arrays were prepared by typing 14 consonants and numerals on a blank index card. The observer recorded his observations by speaking into a tape recorder. The five vowels and the letter *Y* were avoided to make for unambiguous pronunciation. *G*, *Q* and zero were also eliminated because they could be visually confused. The letter *B* was eliminated because it might sound like *P*.

The arrays of the displays were typed



The arrangement devised by Daniel H. Kriegman and Irving Biederman in three rows. The middle row had four figures, the top and bottom rows five each. The blank at the center of the middle row was a point of fixation for the observer. In that spot he saw a point of dim red light made by a small flashlight on the other side of the card. This light was too feeble to illuminate any of the letters on the card.

Six observers took part in the experiments. Each person sat about a foot away from the array and adapted to the dark for about 12 minutes. Then a trial of exposure and recall was done every 30 seconds. The first 10 trials of the session and the first three trials with each filter were for practice. Fifty trials were recorded for each observer. In some sequences the observer began with the dimmest filter and progressed to the brightest. In others the progression was reversed.

The best recalls were made when the light from the flash was brightest. Better recall also resulted from the sequence beginning with the dimmest filter. With the brightest filter the observers were able to recall an average of 12.4 figures out of the 14 in the array. In nine trials at the highest brightness the observers gave perfect reports. (Some of the success may, of course, have been due to increasing familiarity with the figures.)

Apparently accurate and full reporting requires some practice. An observer must keep his gaze steady in order to maintain the afterimage. He must also ignore variations in the illumination of the card that might otherwise interfere with the reading of the figures. With proper concentration observers were able to read the figures in the afterimage even though they did not see them well in the initial flash.

When I repeated this experiment with a similar arrangement, I found that reading the letters in the afterimage was quite strange. Normally one reads by moving one's eyes across the page. With the afterimage I had to gaze steadily to avoid erasing the picture. I do not understand how I then read the letters off to the side. Somehow I thought about what was there and could then recognize the figures. I moved my concentration instead of my eyes.

Bidwell's ghost would still repay investigation. Is the trailing ghost actually related to the snapshot afterimage? What causes the colors in the ghost? Why are the snapshot afterimages colorless? Why are they weakened or erased if your eyes move appreciably? Why are they prolonged if the eyes view a diffuse, flickering field? I would enjoy hearing about what you find to further explain this phenomenon.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

ADVANCED DUNGEONS AND DRAGONS: MONSTER MANUAL. Gary Gygax and Greg Irons. Troubadour Press, 1979.

I CHING: BOOK OF CHANGES. Translated by James Legge. Citadel Press, 1983.

ROG-O-MATIC: A BELLIGERENT EXPERT SYSTEM. Michael Mauldin, Guy Jacobson, Andrew Appel and Leonard Hamey. Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa., 1983.

PREDICTING THE NEXT GREAT EARTHQUAKE IN CALIFORNIA

AN ASSESSMENT OF THE CONSEQUENCES AND PREPARATIONS FOR A CATASTROPHIC CALIFORNIA EARTHQUAKE: FINDINGS AND ACTIONS TAKEN. National Security Council Ad Hoc Committee on Assessment of the Consequences and Preparations for a Major California Earthquake. Federal Emergency Management Agency, January, 1981.

EARTHQUAKE PREDICTION: AN INTERNATIONAL REVIEW. Edited by David W. Simpson and Paul G. Richards. American Geophysical Union, 1981.

FACING GEOLOGIC AND HYDROLOGIC HAZARDS: EARTH-SCIENCE CONSIDERATIONS. Edited by Walter W. Hays. U.S. Geological Survey Professional Paper 1240-B; 1981.

HOW THE TRYPANOSOME CHANGES ITS COAT

NEW GUINEA TAPEWORMS AND JEWISH GRANDMOTHERS: TALES OF PARASITES AND PEOPLE. Robert S. Desowitz in *The Fly That Would Be King.* W. W. Norton and Co., 1981.

BIOCHEMISTRY OF THE VARIANT SURFACE GLYCOPROTEINS OF SALIVARIAN TRYPANOSOMES. M. J. Turner in *Advances in Parasitology,* Vol. 21, pages 69–153; 1982.

BID THE SICKNESS CEASE—DISEASE IN THE HISTORY OF BLACK AFRICA. Oliver Ransford. John Murray, Ltd., 1983.

6 Å-RESOLUTION X-RAY STRUCTURE OF A VARIABLE SURFACE GLYCOPROTEIN FROM *TRYPANOSOMA BRUCEI.* D. M. Freymann, P. Metcalf, M. Turner and D. C. Wiley in *Nature,* Vol. 311, No. 5982, pages 167–169; September 13, 1984.

GAMMA-RAY BURSTERS

PROBABLE OPTICAL COUNTERPART OF A GAMMA-RAY BURSTER. B. E. Schaefer in *Nature,* Vol. 294, No. 5843, pages 722–724; December 24, 1981.

THE GAMMA-RAY BURSTER PUZZLE. Ronald A. Schorn in *Sky and Telescope,* Vol. 63, No. 6, pages 560–562; June, 1982.

HIGH ENERGY TRANSIENTS IN ASTROPHYSICS. Edited by S. E. Woosley. American Institute of Physics Conference Proceedings No. 115, 1984.

BROWNIAN MOTION

ATOMS. Jean Perrin. D. Van Nostrand Company, 1916.

SELECTED PAPERS ON NOISE AND STOCHASTIC PROCESSES. Edited by Nelson Wax. Dover Publications, Inc., 1954.

INVESTIGATIONS ON THE THEORY OF THE BROWNIAN MOVEMENT. Albert Einstein, edited by R. Fürth and translated by A. D. Cowper. Dover Publications, Inc., 1956.

DYNAMICAL THEORIES OF BROWNIAN MOTION. Edward Nelson. Princeton University Press, 1967.

REPRODUCTIVE SUCCESS IN RED DEER

RED DEER: THE BEHAVIOR AND ECOLOGY OF TWO SEXES. T. H. Clutton-Brock, F. E. Guinness and S. D. Albon. University of Chicago Press, 1982.

THE THEORY OF SEX ALLOCATION. E. L. Charnov. Princeton University Press, 1982.

MATERNAL DOMINANCE, BREEDING SUCCESS AND BIRTH SEX RATIOS IN RED DEER. T. H. Clutton-Brock, S. D. Albon and F. E. Guinness in *Nature,* Vol. 308, No. 5957, pages 358–360; March 22, 1984.

PHASED-ARRAY RADARS

RADAR TECHNOLOGY. Edited by Eli Brookner. Artech House, Inc., 1977.

SIDE-LOOKING AIRBORNE RADAR. Homer Jensen, L. C. Graham, Leonard J. Porcello and Emmett N. Leith in *Scientific American,* Vol. 237, No. 4, pages 84–95; October, 1977.

INTRODUCTION TO RADAR SYSTEMS. Merrill I. Skolnik. McGraw-Hill Book Company, 1980.

TRENDS IN RADAR SIGNAL PROCESSING. Eli Brookner in *Microwave Journal,* Vol. 25, No. 10, pages 20–39; October, 1982.

RADAR DETECTION AND TRACKING SYSTEMS. S. A. Hovanessian. Artech House, Inc., 1983.

THE FUNCTIONAL REPLACEMENT OF THE EAR

PLACE AND TIME CODING OF FREQUENCY IN THE PERIPHERAL AUDITORY SYSTEM: SOME PHYSIOLOGICAL PROS AND CONS. E. F. Evans in *Audiology,* Vol. 17, No. 5, pages 369–420; September/October, 1978.

NEURAL PROSTHESES. Gerald E. Loeb, John McHardy, Ellen M. Kelliher and S. Barry Brummer in *Biocompatibility in Clinical Practice: Vol. 2,* edited by David F. Williams, CRC Press, Inc., 1982.

CLINICAL APPLICATIONS OF NEURAL PROSTHETIC TECHNIQUES. F. T. Hambrecht in *Applied Neurophysiology,* Vol. 45, No. 1–2, pages 10–17; January–April, 1982.

COCHLEAR PROSTHESES: AN INTERNATIONAL SYMPOSIUM. Edited by Charles W. Parkins and Samuel W. Anderson in *Annals of the New York Academy of Sciences,* Vol. 405; 1983.

DESIGN AND FABRICATION OF AN EXPERIMENTAL COCHLEAR PROSTHESIS. G. E. Loeb, C. L. Byers, S. J. Rebscher, D. E. Casey, M. M. Fong, R. A. Schindler, R. F. Gray and M. M. Merzenich in *Medical & Biological Engineering & Computing,* Vol. 21, No. 3, pages 241–254; May, 1983.

DAMASCUS STEELS

A HISTORY OF METALLOGRAPHY. Cyril S. Smith. The University of Chicago Press, 1965.

ON THE BULAT—DAMASCUS STEELS REVISITED. Jeffrey Wadsworth and Oleg D. Sherby in *Progress in Materials Science,* Vol. 25, pages 35–68; 1980.

DAMASCUS STEEL-MAKING. Jeffrey Wadsworth and Oleg D. Sherby in *Science,* Vol. 218, No. 4570, pages 328–329; October 22, 1983.

THE AMATEUR SCIENTIST

HOW MANY LETTERS IN BIDWELL'S GHOST? AN INVESTIGATION OF THE UPPER LIMITS OF FULL REPORT FROM A BRIEF VISUAL STIMULUS. Daniel H. Kriegman and Irving Biederman in *Perception & Psychophysics,* Vol. 28, No. 1, pages 82–84; 1980.

WORKSHOPS IN PERCEPTION. R. P. Power, S. Hausfeld and A. Gorta. Routledge & Kegan Paul, 1981.

THE DELAYED ROD AFTERIMAGE. Edward H. Adelson in *Vision Research,* Vol. 22, pages 1313–1328; 1982.

də·'rek·shən

Direction

- *The line or course along which something is moving; e.g., a company approaching $10 billion in sales, a leader in growth markets, zeroed in on the future.*
- *A governing or motivating process; aware management guiding highly resourceful and diverse businesses.*
- *A channel or course of action, like record performance nine straight years.*

Rockwell International is moving fast and smart into the decade ahead. We're leading the way with high technology applied across four diverse business areas.

To keep us moving in the right direction, we've spent $2.2 billion for capital expenditures over the last five years. And we're continuing the pace for '85. To learn more about us, write: Rockwell International, Department 815S-66, 600 Grant Street, Pittsburgh, PA 15219.

Rockwell International

...where science gets down to business

Aerospace/Electronics
Automotive/General Industries


COME ALIVE, COME AND DRIVE
NISSAN
MAJOR MOTION
MOVE OVER!
MAKE WAY FOR THE NISSAN 300 ZX.
FUEL-INJECTED • TURBO-CHARGED • V-6 • 200 HORSEPOWER • 3-WAY ADJUSTABLE SHOCKS.
AWESOME!
AT DATSUN DEALERS
NISSAN
BELT YOURSELF