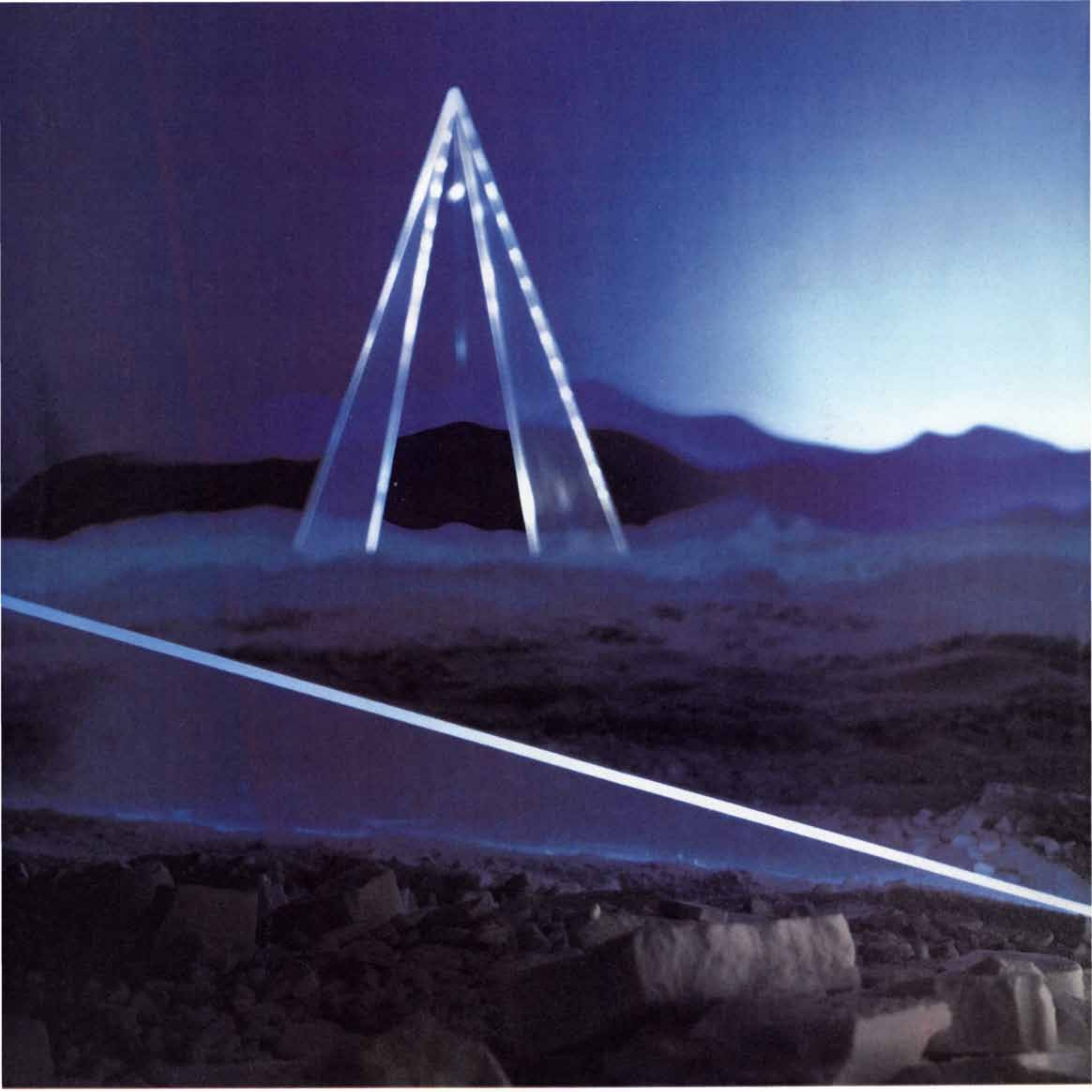
# SCIENTIFIC AMERICAN

ISLAMIC ASTRONOMY

$2.50

*April 1986*



# NOW... THE DAWN OF A NEW

Whenever Zeiss has turned its hand to the very fundamentals of microscopy, microscopy has changed.

In 1846 and the years following, the reputation of the Carl Zeiss workshop was established based upon microscopes of unparalleled mechanical precision, optical quality and reliability.

In 1866 Ernst Abbe, working with Zeiss, formulated his "Theory of Image Formation in the Microscope," which made possible the series production of high-quality microscopes.

In ensuing decades Zeiss has pioneered many of the major advances in microscope production and technique.

Now in 1986 a wider chapter in the history of Zeiss microscopy begins: Zeiss presents microscopes with entirely new infinity color-corrected "ICS Optics" and unique system-integrated "SI Architecture."


Axioplan

# Merkur XR4Ti from Germany. To the question, How well does it perform?

# The sports sedan that asks, How well can you drive?

Its presence alone creates a challenge. Who is this that dares to question the superiority of BMW, Saab and Audi?

It is the Merkur XR4Ti from Germany. And its challenge is rooted firmly in its ability to perform.

XR4Ti takes a sophisticated approach to performance. Its power comes from the advanced technology of blow-through turbocharging and multi-port fuel injection. The result is 175 horsepower (SAE) from a compact 2.3-liter overhead cam 4-cylinder engine. Its redline is an exhilarating 6200 rpm. The suspension is fully independent.

Even XR4Ti's shape enhances its performance by managing the flow of air over, under and around the vehicle. The results are excellent directional stability and improved fuel efficiency.

XR4Ti takes a functional approach to comfort, catering to the driver's needs—whether running down a twisting country road or locked in rush hour traffic.

Merkur XR4Ti was designed and engineered to surpass the demands of serious drivers. It's ready to accept your challenge.

The next move is yours.

For more information, visit a Lincoln-Mercury-Merkur dealer or call **1-800-822-9292.**

**MERKUR XR4Ti**

*Advancing the art of driving.*

Established 1845 **SCIENTIFIC AMERICAN** April 1986 Volume 254 Number 4

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

## THE COVER

The photograph on the cover shows an astrolabe, a two-dimensional brass map of the sky that was widely used by medieval Islamic astronomers (see "Islamic Astronomy," by Owen Gingerich, page 74). The pointers on the open network (the rete) indicate the positions of prominent stars. By rotating the rete about the central pin one could simulate the apparent daily motions of the stars around the north celestial pole; with the help of a celestial coordinate system engraved on an underlying solid plate, one could locate the stars with respect to the horizon and the meridian. Among other things, the astrolabe made it possible to tell time by day or by night. The device was invented by the ancient Greeks, but the oldest dated specimen is the one shown here, made by a Muslim named Nastulus in A.D. 927–28 and now in the Kuwait National Museum. During the Middle Ages, Islamic astronomers preserved and transformed Greek astronomy, and it was from Islam that astronomy—and the astrolabe—later passed to Renaissance Europe.

## THE ILLUSTRATIONS

Cover photograph courtesy of Al-Sabah Collection, Kuwait National Museum

| Page | Source |
|---|---|
| 19 | Thomas Banchoff, David Laidlaw and David Margolis, Brown University |
| 20 | Ilil Arbel (*top*), Ian Worpole (*bottom*) |
| 21–22 | Thomas Banchoff, Hüseyin Koçak, David Laidlaw and David Margolis, Brown University |
| 23 | Ilil Arbel |
| 34–36 | George V. Kelvin, Science Graphics |
| 37 | Sovfoto |
| 38 | Eli Brookner, Raytheon Company |
| 39–40 | George V. Kelvin, Science Graphics |
| 43 | James Kilkelly |
| 44–45 | Ilil Arbel |
| 47–49 | James Kilkelly |
| 50 | Carl Bränden, University of Uppsala (*top*); Sally Black (*bottom*) |
| 51 | Axel Brünger, Harvard University (*top*); Edward Bell (*bottom*) |
| 53–59 | Hank Iken, Walken Graphics |
| 60 | Horst L.Störmer, AT&T Bell Laboratories |
| 75 | Owen Gingerich, Harvard University |
| 76–77 | Hank Iken, Walken Graphics |
| 78 | Thomas Photos, Oxford |
| 79–81 | Hank Iken, Walken Graphics |
| 82 | Bodleian Library, University of Oxford |
| 85–92 | Tom Prentiss |
| 95 | Jon Brenneis |
| 96–98 | Ian Worpole |
| 99 | Jon Brenneis |
| 101 | Charles R. Belinky, Photo Researchers, Inc. |
| 102–106 | Patricia J. Wynne |
| 107 | Thomas Bergman |
| 109 | Don Smithers |
| 110–114 | George Retseck |
| 118–122 | Michael Goodman |

# LETTERS

To the Editors:

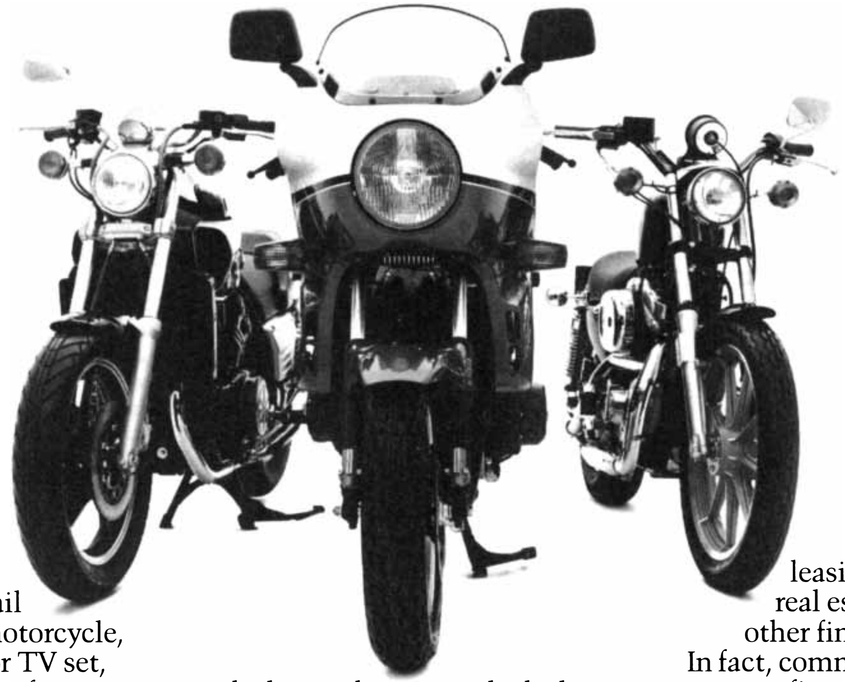
The assertion by Fred H. Tschirley in "Dioxin" [SCIENTIFIC AMERICAN, February] that "concern that this material is harmful to health or the environment may be misplaced" is dangerously premature.

Tschirley repeats the often heard claim that no one has died from dioxin and relies heavily on several epidemiological studies conducted by industry and government agencies. In addition he emphasizes the fact that animal species show a wide range of sensitivity to the acute effects of dioxin and intimates that man is probably less sensitive than the least sensitive of the species that have been studied.

Unfortunately dioxin is a political issue. There are very big economic and political interests that would like the 210 compounds making up the dioxin and furan families declared innocent of any significant damage to humans. There are those in the U.S. Government who are anxious to prove to the world that the massive spraying of jungles in Vietnam with Agent Orange (known to be contaminated with 2,3,7,8-TCDD, the most toxic dioxin congener) caused no damage either to the Vietnamese or to American soldiers. There are others who would like to de-emphasize the need to clean up soil contaminated with dioxin at various toxic-waste dumps. There is also the budding multibillion-dollar refuse-incinerator industry, which is anxious to convince communities across the nation that even though dioxins and furans are formed when unseparated refuse is burned, the levels emitted do not pose a significant health threat. Then there are the chemical companies that manufacture products containing dioxins as contaminants and also various chlorinated plastics such as PVC and polyvinylidine chloride (Saran Wrap); in 1985 Christopher Rappe showed that such plastics form dioxins and furans when they are burned. The chemical companies too have a clear interest in seeing dioxin declared unfit for animals but safe for human beings.

With such clear political and economic interests at stake, we should be very wary of those who would turn off the tap of public concern about dioxin. From the biochemical point of view it would be nothing short of a miracle if a substance so extremely toxic to many animal species proves to be relatively innocuous to humans. We cannot afford to be sanguine about 2,3,7,8-TCDD when a September 1985 U.S.

# One of our fastest growing businesses is other people's business.



When you walk into a retail dealer—for a motorcycle, mobile home or TV set, let's say—think of us.

Because we finance the inventory for over 17,000 retail businesses like these.

The truth is, in a little over a dozen years we've come

from nowhere to being one of the nation's largest sources of inventory financing.

We actually finance more personal computer dealers and motorcycle dealers than anybody else. In the process, ITT Financial has

become a major contributor to our net income.

Not only through inventory financing, but equipment

leasing, commercial real estate loans, and other financial services.

In fact, commercial and consumer financing has become a $6 *billion* business (in receivables) for us. And it's growing over 20% a year. We're concentrating today more and more on businesses like this—businesses that are growing faster than most.

And where the returns are likely to grow just as fast.

Want more about ITT? Phone toll free 1-800-DIAL-ITT for a continuously updated message.

**ITT**

**It's a different world today.**

Environmental Protection Agency report (EPA/600/8–84/014F) indicates it is a carcinogen about 50 million times more potent than vinyl chloride in rats. Results presented in September of 1985 at the Fifth International Symposium on Dioxin in West Germany show that the protein (the Ah receptor) known to bind to dioxin when it enters animal cells is also present in human tissues. Thus humans who have been exposed to dioxin may well be embarked on the first step of the process that in animals leads to toxicity and carcinogenicity. Moreover, it is also known that this Ah receptor is heterogeneous in humans: some people probably have more of the protein than others. We shall therefore need much larger epidemiological studies before we can be certain that the data base is meaningful.

Tschirley and others make much of the fact that animal species vary considerably in their LD50 values (the dose per kilogram of body weight that is necessary to kill 50 percent of the test species). The hamster, for example, is some 5,000 times less sensitive than the guinea pig. Before concluding that man is less sensitive even than the hamster, however, we should note that rhesus macaques (monkeys) have an LD50 value that is intermediate between the values for guinea pigs and hamsters. Moreover, a study by Wilbur P. McNulty and others (1981) indicated that pregnant monkeys are killed by doses of between one and five micrograms per kilogram of body weight, which makes them the most sensitive animals after the guinea pig.

In his article Tschirley suggests that the hamster may be far less sensitive to dioxin's acute effects because the substance is cleared from the hamster body about twice as fast as it is from other tested species. If dioxin's acute toxicity increases as clearance rate decreases, however, human beings are in big trouble: H. Poiger recently showed in an extraordinary experiment (he ate 104 billionths of a gram of 2,3,7,8-TCDD!) that the half-life of dioxin in human tissue could be as long as five years, which is some 60 times longer than it is in the guinea pig and 120 times longer than in the hamster. It is clear there is far more that has to be learned about this troubling substance before we can be certain about anything.

Before we conclude that the dioxins and furans pose little or no health threat to man we need to know the answers to many important questions. What is the exact mechanism by which dioxin exerts its toxic effect on animals? If man appears to be immune to its effects, just how is the mechanism of toxicity circumvented in human tissues? Can the dioxins and furans enhance the effect of other toxic material present in human tissue? While answers to these biochemical questions are pursued, we need epidemiological studies on much larger samples, conducted by workers without the slightest trace of vested interest in the outcome of the work.

With the dioxins and furans the responsible course of action is to err on the side of caution. We must do our utmost to ensure that the smallest possible amounts of dioxins and furans enter our environment. If the result is to slow down certain industries a little, so be it: history will not blame us for being too cautious, but it will have harsh things to say about those who call our concern misplaced before we have full answers to key questions about these troubling substances.

PAUL CONNETT

Department of Chemistry
St. Lawrence University
Canton, N.Y.

To the Editors:

Dr. Connett presents interesting points, some of them valid, in my view, but he appears to have missed the central thrust of my paper, namely a recognition that 2,3,7,8-TCDD is indeed a highly toxic chemical and there is much we do not know about its mechanism of action, but that there is now a substantial body of data demonstrating that humans are much less sensitive to TCDD than was feared on the basis of tests with experimental animals. Such information is critical for informed decision making. The public has a right to know: it should be told of good news as well as bad news.

While I agree that the evidence from human exposures does not allow a definitive statement about whether TCDD is or is not a human carcinogen, the evidence that is available should not be ignored. The same is true for other possible chronic effects. We live in a pragmatic world, ever beset by the need to make decisions in the face of incomplete data and to establish priorities for allocation of limited resources. In that context an undue allocation to research on TCDD may preclude the development of research data on other materials that appear to have a greater potential for adverse effects on human health and the environment. Dibenzofurans, benzene and other organic solvents to which human exposure is high and naturally occurring carcinogens contained in foods are cases in point.

I am disturbed by Connett's reference to all dioxins and furans, when only 2,3,7,8-TCDD was discussed in my paper, and by his broad-brush treatment of the political nature of the dioxin issue, with the strong implication that private enterprise and government are guided only by considerations of self-interest. His treatment of both issues is simplistic and tends to becloud rather than clarify.

I did not declare "innocent of any significant damage to humans" the 210 compounds that make up the dioxin and furan families. In fact, aside from saying that there are many chlorinated dioxins, I discussed 2,3,7,8-TCDD only and did not even mention the furan compounds.

The compound 2,3,7,8-TCDD is indeed a political issue, but it has more dimensions than the "economic and political interests" Connett mentions. If I were to accept the axiom he espouses, I would have to be wary of the research of Richard J. Kociba and his colleagues (Dow Chemical Company toxicologists), who demonstrated the carcinogenicity of TCDD in rats; of Dow analytical chemists who have made major contributions to the detection of dioxin at levels of parts per trillion and to the separation of the 75 chlorinated dioxin isomers; of Alvin L. Young and his colleagues (Air Force personnel) who have conducted major studies on ecological effects; of George D. Lathrop et al. (Air Force workers) who have investigated mortality and morbidity in the pilots and crewmen who took part in the spraying of Agent Orange, and so on. Should I also be wary of research by university scientists whose ability to obtain additional research grants may be enhanced when they publish evidence of adverse health or environmental effects? I would hope not! The point is that published scientific research stands or falls on its merits, not on the affiliation of the investigator. Guilt by association should have no place in rational decision making.

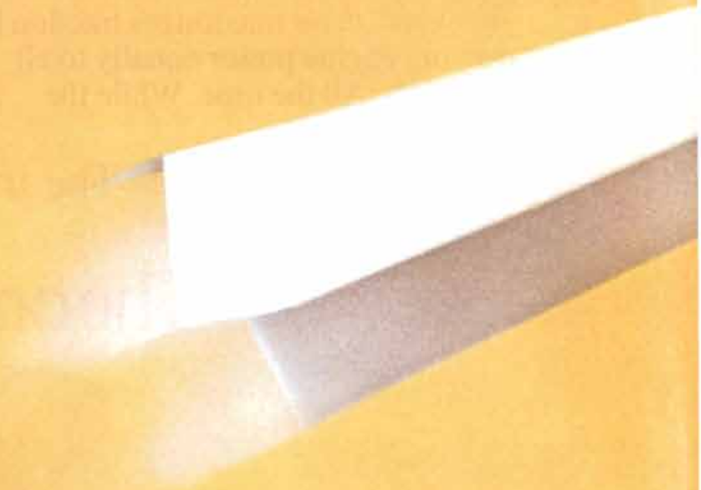
I agree that there is much that is not known about 2,3,7,8-TCDD. Some of the gaps will be filled by research now under way; others may never be filled. We should not adopt a casual attitude toward the addition of more TCDD to the environment, but neither should we perpetuate an unfounded fear of health effects when the available evidence does not support it. There are too many other pressing problems to allow an overly lavish allocation of resources for dioxin research.

FRED H. TSCHIRLEY

Williamston, Mich.

**Avionics.**

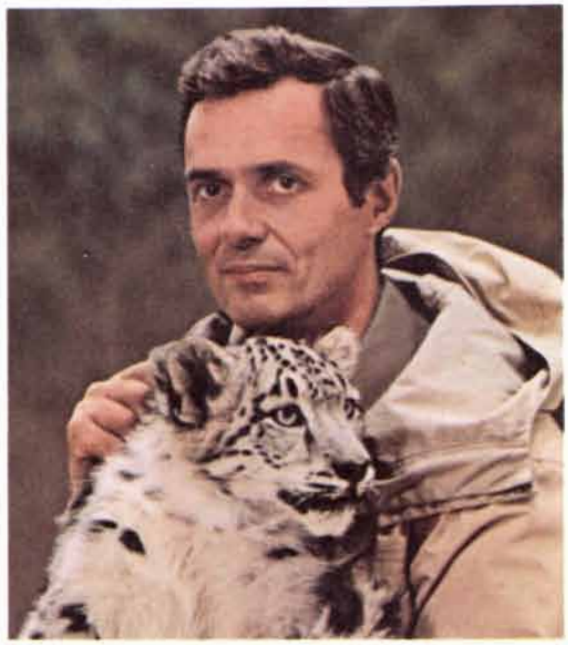
It's just taking off. And the future belongs to those who provide new answers to old questions—to Eaton.



EATON

**Growing into the future.**

**Automotive • Electronics • Defense • Capital Goods**



Snow leopards in the Himalayas. Tigers in the wilds of India. Mountain gorillas and lions in Africa. Jaguars in the swamps of Brazil.

George Schaller, pictured here with a snow leopard, has spent years in remote and rugged places studying the natural history of rare animals—and fighting for their survival.

He sees these animals as symbols of the habitats in which they live. Preserve their habitats and thousands of other plants and animals will be assured of a home.

As director of Wildlife Conservation International, a division of the New York Zoological Society, Schaller and the staff have helped establish more than 50 reserves around the world.

*One of the 1000 remaining giant pandas.*

He points out that the destruction of environments is now so drastic that, in the decades ahead, the nature of life on earth will be irrevocably changed.

For Schaller, saving fragments of nature is an urgent task.

Recently he faced one of his greatest challenges. With Chinese scientists, he collaborated in a project to save the 1000 giant pandas still alive in the wild. Currently he is working on the Tibetan plateau to help preserve the wildlife of those remote uplands.

# "Future generations must be inheritors... not just survivors."

**George Schaller**

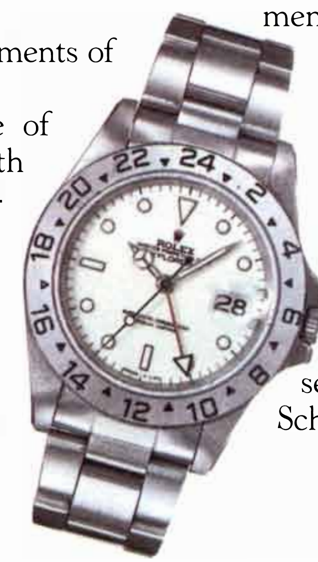
Since his work takes him to some of the most forbidding places on earth, choosing the right equipment is crucial.


*China: Site of panda rescue mission.*

Not surprisingly, Schaller wears a Rolex.

"My watch must be absolutely reliable, as animal observations are recorded under the most demanding conditions. My Rolex has never let me down."

Inhospitable conditions seem to pose no problem for George Schaller. Or his Rolex.



**ROLEX**

*The Rolex Explorer II Oyster Perpetual Chronometer in stainless steel.*
*Write for brochure. Rolex Watch U.S.A., Inc., Dept.004, Rolex Building, 665 Fifth Avenue, New York, New York 10022-5383.*
*World headquarters in Geneva. Other offices in Canada and major countries around the world.*

# 50 AND 100 YEARS AGO

## SCIENTIFIC AMERICAN

APRIL, 1936: "New York to London in 36 hours, with passengers, mail and express! A decade ago this might well have been the title for a book by a contemporary Jules Verne. Today it is an imminent reality. The North Atlantic, scene of a score of heroic pioneer flights following that first solo blazing of its trail by Colonel Charles A. Lindbergh in May, 1927, which may be said to have touched off the fuse of air transport development the world over, is to become, almost certainly this year, the field of scheduled operation. The problem has been not the perfection of flying equipment fully equal to the task but the perfection of international agreements."

"C. G. Abbot, the head of the Smithsonian Institution, has changed the design and raised the efficiency of his solar heat concentrator. There is now some promise that, with a few more such increases, solar heat, as an ordinary power source, may enter the market in competition with the one source of energy now most generally efficient, that is, coal."

"The second annual inventory of migratory waterfowl spending the winter in the United States, which was completed in January, was extraordinary in the completeness of its organization. The inventory, done under the direction of the Bureau of Biological Survey, was made after long planning and with the largest number of qualified observers ever assembled. When all reports have been analyzed, the Bureau hopes to have available a reasonably accurate estimate of the continental supply of waterfowl."

"Despite the fact that people are more map-conscious than ever before, not more than 47 percent of the United States has been completely mapped to show elevations and slopes. Dr. William Bowie of the Coast and Geodetic Survey says that the United States is one of the most backward of all the organized countries of the world."

"The title of this note [$44^2$] is simply another way of writing the year 1936, for this number is an exact square of 44. This is the first time since 1849 that the year has resulted from a perfect square, and it will not happen again until A.D. 2025."

## SCIENTIFIC AMERICAN

APRIL, 1886: "The power of flying, being denied to man, has always been one of the objects most desired by him, although hitherto he has not succeeded in attaining it. If there were any large birds feeding on grains and possessing strong flying powers, they would no doubt have been domesticated long since, and made subservient to man's use, like horses and other animals. But unfortunately all large birds possessing strong wing power are carnivorous and untamable, and so we shall have to rest content with terrestrial locomotion till we have succeeded in solving the mechanical problem of propelling and steering balloons. We are still a long distance from this result, and it is at least very doubtful whether it will ever be attained."

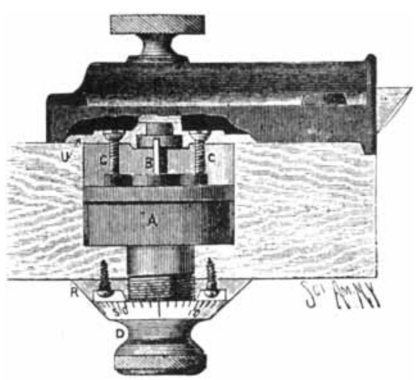
"A new sweetening agent has been produced from coal tar. It is known to chemists as benzoyl sulphuric imide, but it is proposed to name it saccharine. The discoverer is Dr. Fahlberg, and its preparation and properties were recently described by Mr. Ivan Levinstein at a meeting of the Manchester section of the Society of Chemical Industry. Saccharine is about 230 times sweeter than the best cane or beet-root sugar."

"Although water usually freezes at 32 degrees Fahrenheit, and ice melts when above that point, the result is not uniform in either case. If water, for instance, be kept in a clean, smooth-sided vessel, and perfectly still, it is possible to keep it from freezing until it reaches a temperature of 15 degrees. With regard to the melting point of ice, the temperature is more uniform, since the solid ice is not subject to the law of motion as water is, but there are ways of precipitating the melting of ice. Thus, for instance, if a block of ice be subjected to a heavy pressure, the melting point can be reduced to 18 degrees F."

"The 'Bad Lands' of Dakota are said to owe their origin to the burning of the coal deposits that once existed there. The lands are situated principally along the Cheyenne and Grand rivers and the Little Missouri. In the long ago the valleys of these streams must have been filled with drift wood. Then followed a period of drift, which buried the accumulation of wood under two or three hundred feet of sediment, sand and gravel. The buried wood in time became coal. Either from spontaneous generation or from electricity, fires were started in those veins, and they gradually burned out, restoring in part the old water courses. Looking upon them, here you see patches of slag, there great bowlders, showing unmistakable evidences of great heat, and on every hand scoria or burned clay, resembling broken brick."

"The remains of a gigantic extinct sea turtle have been found near Fort Wallace in western Kansas. The question may arise: How did the sea turtle become buried in a bluff in the state of Kansas? A natural supposition would be that Kansas is the bed of a former ocean, and so it is. Ages ago, in what is called by geologists the Cretaceous period, that part of the world was the bed of a great sea, in which the great turtle swam. Gradually the crust of the earth was raised, the water fell back or became inclosed and left the inhabitants of the Cretaceous sea high and dry."

"Our usual expression for security is that we have placed valuables 'under lock and key,' but as the lock may be picked and the key may be lost, this does not always describe the best fastenings. In some of the improved 'Champion' locks there is neither key nor key hole. Doors provided with them may be opened from either side, the 'Open Sesame' being a knowledge of the combination of figures by which the knob may be made to turn and the door open. Perhaps it may be feared that the combination might be forgotten, but it must be remembered that a key is not only liable to be left behind, but as well to be lost or duplicated. The combination necessary for the unlocking of a keyless lock may be recorded in any number of places, and it can be recorded in such a way that detection would be impossible."




*The keyless door lock*

# THE AUTHORS

MIROSLAV NINCIC ("Can the U.S. Trust the U.S.S.R.?") is associate professor of politics at New York University. He holds a degree in political science and diplomacy (1970) from the Free University of Brussels and an M.A. in international relations (1974) and a Ph.D. in political science (1977) from Yale University. Before accepting his current position he was associate professor of political science at the University of Michigan. Nincic has written several books on the arms race.

MARTIN KARPLUS and J. ANDREW McCAMMON ("The Dynamics of Proteins") are respectively Theodore William Richards Professor of Chemistry at Harvard University and M. D. Anderson Professor of Chemistry at the University of Houston. A native of Austria, Karplus got his B.A. at Harvard in chemistry and physics in 1951 and then studied physical chemistry with Linus Pauling at the California Institute of Technology, where he received his Ph.D. in 1954. He spent two years at the University of Oxford working with Charles A. Coulson and then joined the faculty at the University of Illinois at Urbana-Champaign. Karplus moved to Columbia University in 1960 and to his present post at Harvard in 1965. McCammon studied chemistry and physics at Pomona College and Harvard, where he obtained his doctorate in 1976 with Karplus and John M. Deutch. McCammon did postdoctoral work at Harvard for two years before going to the University of Houston.

BERTRAND I. HALPERIN ("The Quantized Hall Effect") is professor of physics at Harvard University. His undergraduate training was at Harvard, where he got his A.B. in 1961. He went on to study physics at the University of California at Berkeley, earning a Ph.D. there in 1965. During the year after he finished his graduate work he was a National Science Foundation Postdoctoral Fellow at the University of Paris. He subsequently worked at Bell Laboratories in Murray Hill, N.J. From 1973 to 1980 he served as associate editor of *Reviews of Modern Physics.* In 1982 Halperin received the American Physical Society's Oliver E. Buckley Condensed Matter Physics Prize.

OWEN GINGERICH ("Islamic Astronomy") is an astrophysicist at the Smithsonian Astrophysical Observatory in Cambridge and professor of astronomy and the history of science at Harvard University. He got his undergraduate degree at Goshen College in 1951 and his M.A. and Ph.D. from Harvard in 1953 and 1962 respectively. From 1955 to 1958 he taught astronomy and physics at the American University of Beirut.

EBERHARD GWINNER ("Internal Rhythms in Bird Migration") is head of the Vogelwarte Radolfzell, which is an ornithological station of the Max Planck Institute for Behavioral Physiology, and an adjunct professor of zoology at the University of Munich. He had his undergraduate training at the University of Tübingen and got his Ph.D. there in 1964 for work on the social behavior of the raven. Subsequently he worked on various aspects of biological rhythms in bird behavior at the institute headed by Jürgen Aschoff in Erling-Andechs, the University of Washington and Stanford University. Although Gwinner now works mainly in the laboratory, he still considers himself a field biologist and spends as much time as possible outdoors, particularly in the Alps.

DANA Z. ANDERSON ("Optical Gyroscopes") is assistant professor of physics at the University of Colorado at Boulder and a fellow of the Joint Institute for Laboratory Astrophysics there. He earned a B.S. in electrical engineering from Cornell University in 1975 and a Ph.D. in physics from the University of Arizona in 1981. As a postdoctoral fellow at the California Institute of Technology he worked on the development of an interferometer for detecting gravitational waves. Anderson's concern about the gap between science and the humanities led him to collaborate with Eric White of the English department at the University of Colorado in the creation of a course titled "Contemplating Nature: The Human Implications of Modern Physics."

HUGO LAGERCRANTZ and THEODORE A. SLOTKIN ("The 'Stress' of Being Born") share an interest in the development of the newborn. Lagercrantz is associate professor of physiology and perinatal medicine at the Karolinska Institute in Stockholm, where he received his Ph.D. in physiology under the tutelage of the late Ulf von Euler. Since 1976 he has mainly studied the release and function of catecholamines (which function both as hormones and as neurotransmitters) in newborn infants. Slotkin is professor of pharmacology and psychiatry at the Duke University School of Medicine. He has a B.S. in chemistry (1967) from Brooklyn College and a Ph.D. in pharmacology and toxicology (1970) from the University of Rochester. Over the past decade he has worked extensively on the biochemistry, physiology and pharmacology of the developing nervous system and on the regulation of cardiovascular function in the newborn. In 1976 Slotkin was named the outstanding young investigator of the North Carolina Heart Association, and in 1982 he was the recipient of the John Jacob Abel Award from the American Society for Pharmacology and Experimental Therapeutics.

DON SMITHERS, KLAUS WOGRAM and JOHN BOWSHER ("Playing the Baroque Trumpet") bring diverse skills to their investigation of the trumpet. Smithers is a lecturer on the history of music and musical instruments and a member of the Royal Musical Association. He studied music at Hofstra University, getting a B.S. in 1957. The University of Oxford awarded him a Ph.D. in the history of music in 1967. He has held positions at Syracuse University and the Royal Dutch Conservatory of Music. It is Smithers who plays the baroque trumpet fanfare heard as the introduction to the CBS television news program "Sunday Morning" with Charles Kuralt. Wogram is in the division of physical acoustics at the German Bureau of Standards. He received his degree in electrical engineering from the Technical University of Braunschweig in 1967 and his doctoral degree in physical acoustics there in 1972. He is now investigating the acoustics of wind and keyboard instruments. A practicing musician, Wogram plays the trombone and the baritone horn. Bowsher is a lecturer in acoustics at the University of Surrey and vice-president of the Institute of Acoustics. He holds an undergraduate degree and a Ph.D. from the Imperial College of Science and Technology in London (1957). Before taking his current position he was a senior scientific officer in the acoustics section of the National Physical Laboratory in Middlesex. A former professional bass trombonist, Bowsher now plays for recreation.

SALVADOR E. LURIA, who reviews *The Transforming Principle: Discovering that Genes Are Made of DNA,* by Maclyn McCarty, is Institute Professor Emeritus at the Massachusetts Institute of Technology. He shared the Nobel prize in physiology or medicine in 1969.

# COMPUTER RECREATIONS

## *A program for rotating hypercubes induces four-dimensional dementia*

by A. K. Dewdney

"My husband has disappeared into thin air, and I think you had something to do with it!" The woman on the telephone was Cheryl, and she was clearly upset. Her husband Magi, my microcomputer amanuensis at the University of Western Ontario, had apparently vanished while viewing a computer program I had suggested he write. The program rotates a four-dimensional analogue of a cube called a hypercube and projects it on a display screen. Cheryl went on in agitation: "There's a weird pattern of lines on the monitor and his clothes are lying in a heap near the chair. He must have been wearing these strange colored glasses made out of cardboard. And look at this—his socks are still in his shoes!"

Here, it seemed to me, was an obvious case of four-dimensional dementia. Victims become convinced they have stepped out of ordinary space and entered a higher-dimensional reality invisible to others. The delusion that one has disappeared can be so powerful that others take part in it: the victim can enter a room full of people and seem invisible to all. Fortunately Magi's case has a happy ending; I shall save it for last. In the meantime I submit the hypercube program to the wider public with what I hope is a responsible warning: Readers likely to fall prey to Magi's dementia are urged not to write the program or to view its output on a display screen. Potential victims include anyone with a history of obsession about the higher dimensions or anyone who is even occasionally tempted by the prospect of unknown realities.

The fourth dimension has been a vehicle for physical and metaphysical speculation at least since the 19th century. The idea of a fourth, physical dimension culminated in Einstein's theories of special and general relativity; space and time together make up a four-dimensional continuum in which all real events are timelessly frozen. This view of the universe may be undergoing dimensional modifications; the so-called Kaluza-Klein theories introduce seven or more new dimensions in the form of miniature hyperbubbles attached to every point of spacetime [see "The Hidden Dimensions of Spacetime," by Daniel Z. Freedman and Peter van Nieuwenhuizen; SCIENTIFIC AMERICAN, March, 1985].

The fourth dimension that I have come to know and love is the child of mathematics. Readers in ordinary rooms have a three-dimensional coordinate system suspended overhead. Three walls meet in each corner of the room, and from that corner radiate three lines, each of which is the meeting place of a pair of walls. Each line is perpendicular to the other two lines. Can the reader imagine a fourth line that is perpendicular to all three lines? Probably not, but that is what mathematicians require in setting up the purely mental construct called four-dimensional space. You now have the chance to explore this space in a personal way and without danger to your person. You have only to write the program I call HYPERCUBE.

HYPERCUBE can trace its origins to a film produced in the mid-1960's by A. Michael Noll, then at Bell Laboratories, that depicts the two-dimensional shadows of four-dimensional objects moving in four-dimensional hyperspace. The program as it now stands, however, was developed by Thomas Banchoff and his colleagues in the Computer Graphics Laboratory at Brown University, and my inspiration for this column comes from the fascinating images it generates [*see illustrations on pages 19, 21 and 22*]. Banchoff, who is a professor of mathematics, directs the visual exploration of higher-dimensional surfaces and spaces as a complement to his writing and research as a geometer. In 1978 he and Charles Strauss produced a 9½-minute computer-generated color film that has since become a classic in the mathematical underground: *The Hypercube: Projections and Slicing.* (The film can be obtained from the International Film Bureau, Inc., 332 South Michigan Avenue, Chicago, Ill. 60604.) Banchoff is also probably the leading expert on the life and work of Edwin A. Abbott, the English clergyman and teacher who in 1884 wrote *Flatland,* a tale of imagined life in two dimensions.

Banchoff and his colleagues have devised striking images that illustrate properties of four-dimensional objects. The images on page 19, for example, depict the rotation of a four-dimensional hypercube in four-dimensional space. To appreciate the images consider the shadow cast by an ordinary cube on a plane: the shadow can resemble a square inside a square. If the appropriate faces of the cube are shaded, the shadow is a square with a square hole in it [*see bottom illustration on page 20*].

Similarly, when a hypercube is illuminated from a point "above" ordinary space in the fourth dimension, the three-dimensional "shadow" cast by the hypercube can resemble a cube inside a cube. The inner cube is surrounded by six six-sided polyhedrons that can be regarded as distorted cubes. The four distorted cubes adjacent to the sides of the inner cube fit together to form the solid figure whose surface is the boxlike torus shown in Banoff's images. The other two distorted cubes, the inner cube and the outer cube also form a solid torus, which is not shown. As the hypercube rotates, the square hole in the visible torus seems to move toward the viewer. Those who write the program HYPERCUBE will see similar changes, albeit not so realistic or continuous.

The images on pages 21 and 22 are from a forthcoming film by Banchoff and his colleagues Hüseyin Koçak, David Laidlaw and David Margolis: *The Hypersphere: Foliation and Projections.* The hypersphere is a far more complex object than the hypercube, and I shall not describe it in detail. Nevertheless, one can begin to appreciate the images by considering an ordinary sphere. If the sphere is initially at rest on a plane tangent to its south pole and a light is fixed at the initial position of its north pole, the shadow cast on the plane by the lines of latitude is a series of concentric circles [*see bottom illustration on page 20*]. If the sphere is rotated while the light is kept fixed, the images of the circles may become nonconcentric, and the image of any circle that passes through the source of light is a straight line.

# What will be.

It will be new. All that has been will be the beginning. It will be a special shape. One that will cut cleanly through the air.

It will be lower to the ground. It will be larger inside. It will be a joy to drive. It will be the exciting Honda Accord LXi Hatchback.

The hood angles down for greater aerodynamic efficiency. Under the hood a larger 2-liter engine designed with Honda Programmed Fuel Injection delivers smooth power.

A unique suspension system provides outstanding handling. It's called double wishbone. Race cars have used it for years. Nothing uses it like the Accord does.

Driving one will be an experience you will want to repeat. For once, the future is yours to see. HONDA

The New Accord

# by more people can by more people.

And it's the fastest way to network those users, too. Because now, true multi-user capabilities for local area networking are built right in.

dBASE III PLUS can also help put developers in the fast lane. With a new Data Catalog and more than 50 new commands and functions. Plus code encryption and linking, improved debugging aids, assembly language calls and much more.

For the name of the Ashton-Tate® dealer nearest you,* call 800-437-4329, Extension 2844.**

And get your hands on dBASE III PLUS. It's the software more people can look forward to using.

*Upgrades are available to all dBASE III owners. **In Colorado call (303) 799-4900, Extension 2844. Requires IBM® PC or 100% compatible. Trademarks/owners: Ashton-Tate, dBASE III PLUS/Ashton-Tate; IBM/International Business Machines Corp.





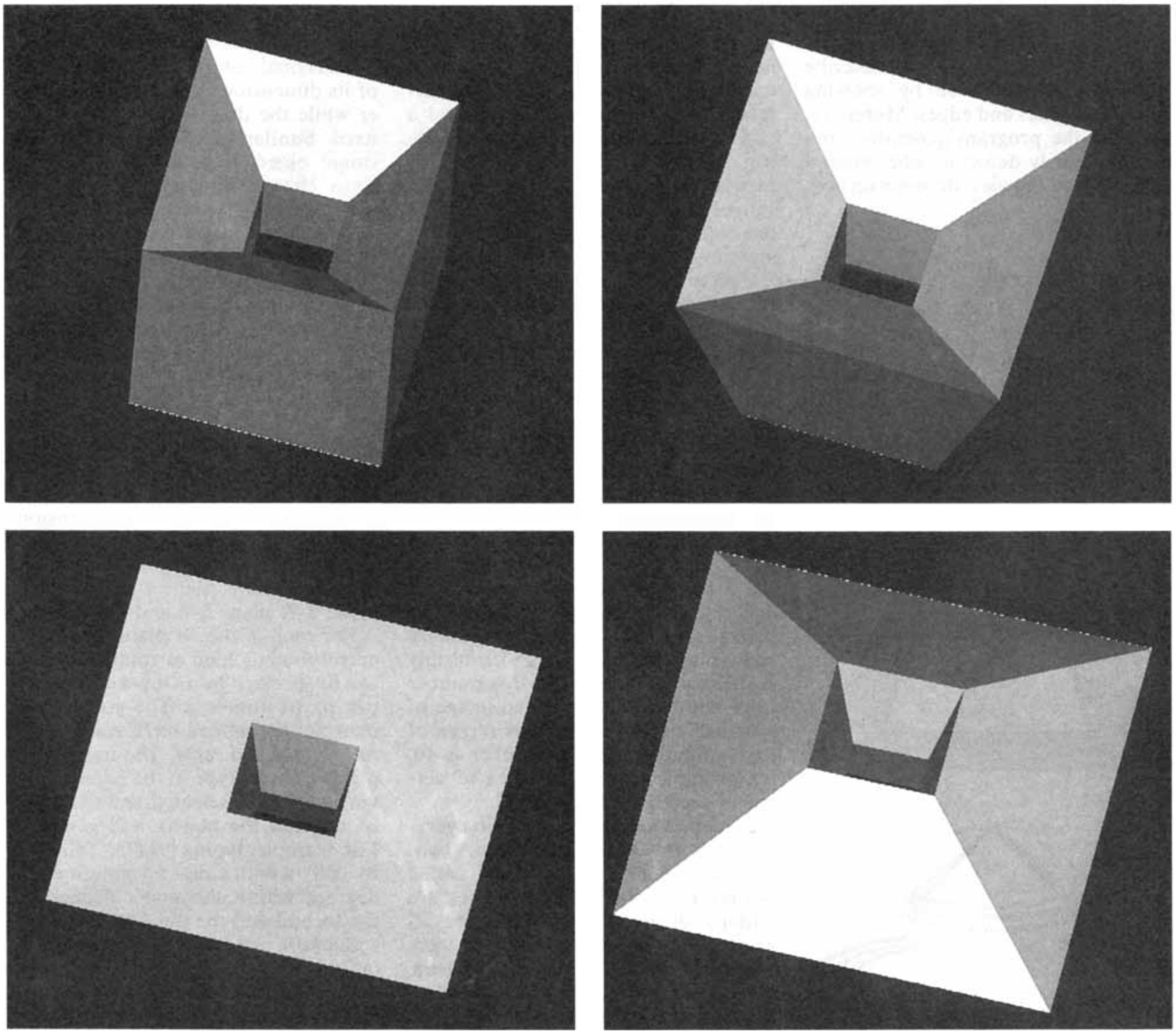
Similarly, the three-dimensional "shadow" cast by a hypersphere can be viewed as a series of concentric toruses [*see illustration on page 21*]. The toruses are made more readily visible in Banchoff's images by cutting away parts of one torus along strips that wind around it. When the hypersphere is rotated, the toruses appear to swell up and sweep past one another. Any torus that passes through the source of light becomes infinitely large [*see illustration on page 22*].

Dimensional analogies are valuable tools in constructing and understanding four-dimensional phenomena. The hypercube, for example, is derived from the cube just as the cube is derived from the square. To get the cube from the square lift the square in a direction perpendicular to its plane, up to a height equal to its side [*see top illustration on next page*]. The new cube has eight vertexes, twice as many as the initial square, and 12 edges, four from the initial square, four from the final square that is lifted away from the initial square and four that arise when vertexes in the initial square are connected to their counterparts in the final square. The cube also has six square faces, one coincident with the initial square, one coincident with the final square and one erected between each of the four pairs of edges that make up the initial and final squares.

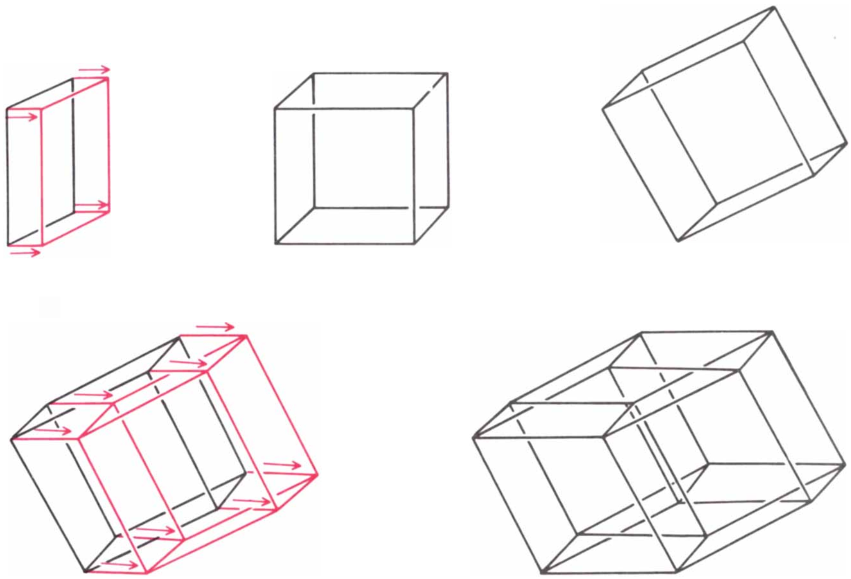
If one pretends for the moment that an additional dimension is available, the same operation can be repeated with the cube: "lift" the cube away from ordinary space in the direction of the extra dimension, out to a distance equal to the side of the cube [*see top illustration on next page*]. The result is a hypercube. But in what direction does the extra dimension lie? I cannot explain that. Even a photograph of me pointing into the fourth dimension would be utterly useless. My arm would simply appear to be missing.

Nevertheless, the number of vertexes, edges, faces and hyperfaces (ordinary cubes) that make up the hypercube can readily be counted. The number of vertexes is just the number of vertexes in the initial cube plus the number in the final cube, or 16. Each of the eight vertexes in the initial cube is joined by an edge to one of the eight vertexes in the final cube, and there are also 12 edges in each of the two cubes. Hence there are 8 + 12 + 12, or 32, edges in the hypercube. One can also show that the hypercube has 24 ordinary faces and eight hyperfaces.

I am indebted to David Laidlaw for

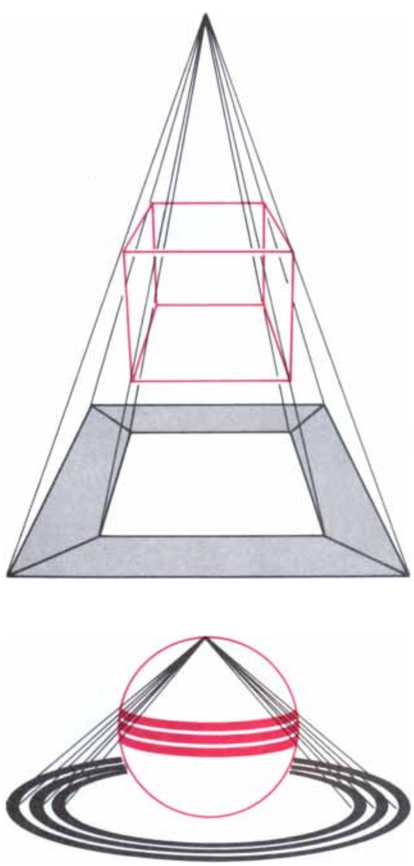


*Rotation of a four-dimensional hypercube through dimensions 2 and 4, projected into ordinary three-dimensional space*



*How a plane generates a cube and a cube generates a hypercube*

an explanation of HYPERCUBE. The version of the program I shall describe represents a hypercube by showing only its vertexes and edges. Moreover, the view the program generates does not necessarily depict a cube inside a cube; instead the view depends on how



*Projections of the cube and the sphere*

HYPERCUBE is implemented and on how it is run. Every time the hypercube in the program is rotated the vertexes swing into new positions and a new, oddly confusing view of the object results. With continued experimentation, however, the views begin to make a strange kind of sense, and one feels on the threshold of something awesomely spacious and inviting.

The 16 vertexes of the hypercube in the program are numbered from 0 to 15 according to a simple scheme. If each number is rewritten in binary form and converted into an array of four bits, a miniature coordinate system emerges. The binary digits of 13, for example, are 1 (that is, one 8), 1 (one 4), 0 (zero 2's) and 1 (one 1). The binary number can then be written as the array (1,1,0,1), which almost gives a practical coordinate system for the initial position of the hypercube. (It is not a position that resembles a cube inside another cube.) To convert the binary array into useful coordinates, change the 0's to −1's and multiply each member of the array by a number large enough to generate an image of practical size on the display screen of the computer. If the multiplier is 10, for example, the coordinates of vertex 13 are (10,10,−10,10).

Dimensions seem to creep in everywhere as HYPERCUBE is written. A two-dimensional matrix, or array, called *vert* preserves the vertexes as they are initially defined. Since there are 16 vertexes with four coordinates each, *vert* is a 16-by-4 matrix of 64 numbers; *vert*(*i,j*) is the *j*th coordinate of the *i*th vertex. The program HYPERCUBE holds the matrix *vert* inviolate; *vert* is defined at the beginning of the program and its contents are then transferred to a second 16-by-4 matrix called *cube.* The matrix *cube* can be thought of as a working matrix; its contents are continually altered by the rotations carried out in the program.

HYPERCUBE is divided into three major sections following the initialization of *vert:* the selection of the desired rotation of the hypercube, the calculation of the coordinates of the rotated hypercube and the display of the result on the monitor. If the rotating object were three-dimensional, one could select the rotation by specifying the orientation of the axis of rotation and the angle of the rotation about the axis. For a rotating four-dimensional object, however, picking an axis of rotation does not determine a rotating plane: remember that there are two nonequivalent directions perpendicular to a given plane. On the other hand, even in four-dimensional hyperspace it remains true, as it does in ordinary space, that a rotation can affect just two dimensions at a time. If a three-dimensional object is rotated, two of its dimensions swing into each other while the third dimension remains fixed. Similarly, when a four-dimensional object is rotated, two dimensions change direction in the space while the other two remain fixed.

There are many ways a four-dimensional object can be rotated to a new position. It turns out, however, that any position can be reached by applying a sequence of rotations limited to motions within the planes defined by the coordinate axes of the surrounding four-dimensional space. There are four coordinate axes in a four-dimensional space, numbered, say, from 1 to 4, and there are six ways any two of them can be combined. Hence there are six planes within a four-dimensional space determined by the coordinate axes: plane 1–2, the plane determined by axes 1 and 2, plane 1–3, plane 1–4, plane 2–3, plane 2–4 and plane 3–4.

For each of the six planes there is a corresponding kind of rotation, which can be specified by a 4-by-4 square matrix of 16 numbers. The six rotation matrixes are named *rot12, rot13, rot14, rot23, rot24* and *rot34.* The user of the program must type in the name of the kind of matrix selected and the angle of rotation the matrix will generate. For example, typing "rot23" followed by "60" would cause a rotation of 60 degrees within the plane defined by the second and the third axes.

Suppose one wants to confine the rotation of the hypercube to the third and fourth dimensions, the most mysterious rotation of all. The rotation matrix *rot34* is applied. Its entries are 0's, 1's and three other numbers distributed according to the following pattern:

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & -b & a \end{matrix}$$

The angle of the desired rotation in degrees is selected and stored in the variable *ang,* and the numbers *a* and *b* depend on *ang: a* is equal to *cos*(*ang*) and *b* is equal to *sin*(*ang*), where *cos* and *sin* are the trigonometric functions sine and cosine.

The rule for generating the other five rotation matrixes is simple. The *a*'s appear on the main diagonal of each matrix in positions that correspond to the dimensions affected by the rotation. The *b*'s appear at all the other intersections of rows and columns that correspond to the rotating dimensions. All other entries on the main diagonal are 1's, and the rest of the entries in the matrix are 0's. For example, *rot13* is the following matrix:

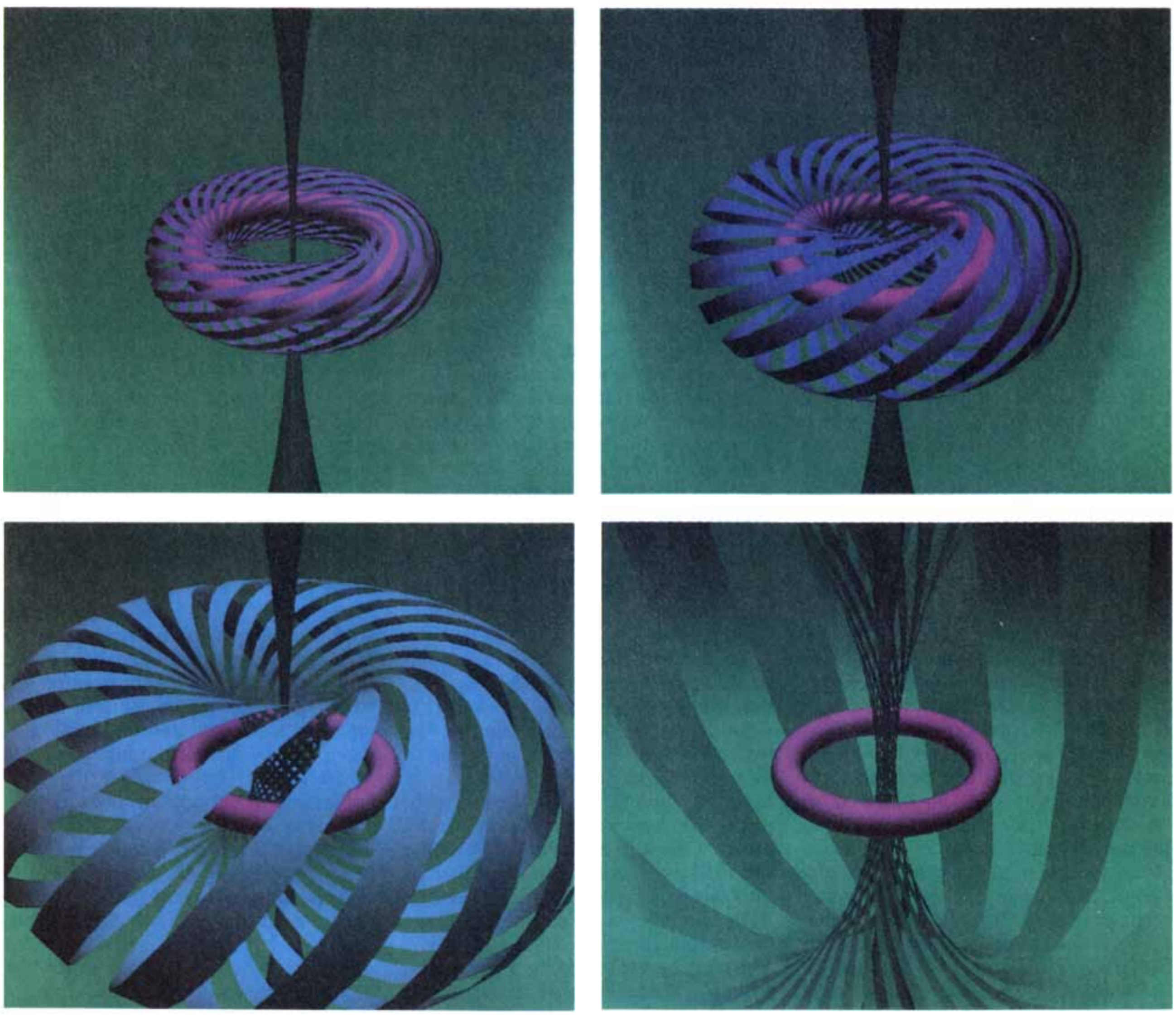
$$\begin{matrix} a & 0 & b & 0 \\ 0 & 1 & 0 & 0 \\ -b & 0 & a & 0 \\ 0 & 0 & 0 & 1 \end{matrix}$$

When the desired rotation matrix has been selected, it is assigned by HYPERCUBE to a special matrix, *rote.* The assignment can be made conveniently by employing a double loop, the inner loop for the sequence of numbers across a row of the matrix and the outer loop for the sequence of rows.

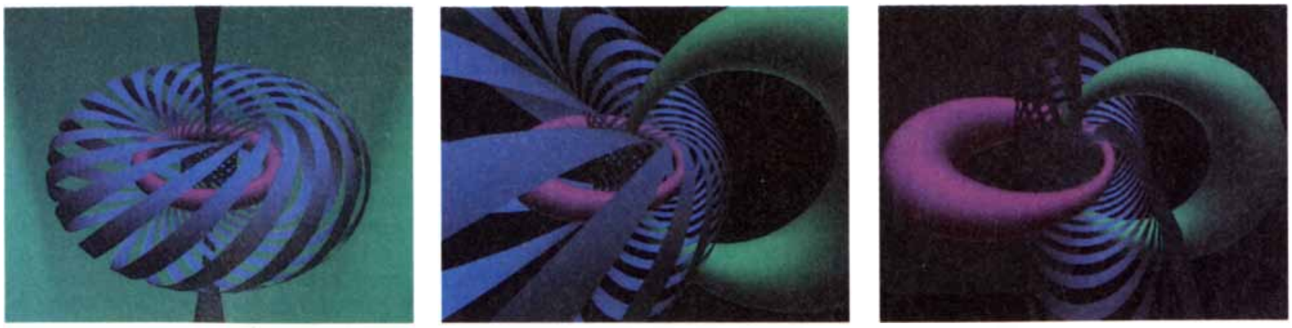
The calculation of the coordinates of the rotated hypercube is done by "multiplying" the matrix *cube* by the selected rotation matrix *rote.* The product of the two matrixes is stored temporarily in a third matrix called *temp,* and it gives the coordinates of the rotated hypercube. That product is found according to the rules of matrix multiplication, a standard operation on matrixes of numbers. It results from an orderly orgy of multiplications embedded in three nested loops.

*Temp,* like *vert* and *cube,* is a 16-by-4 matrix of 64 numbers: it gives four coordinates for each of the 16 vertexes of the rotated hypercube. Each number *temp*(*i,j*) in the matrix is designated by a pair of indexes *i* and *j*. For example, *temp*(13,3) is the third coordinate of the 13th vertex of the rotated hypercube. Its value is the sum of four products of numbers drawn in a precisely defined way from the matrix *cube* and from the matrix *rote* [*see illustration on page 23*]. The innermost loop in the program can therefore be indexed by the letter *k,* which runs from 1 to 4, and that loop returns the value of one entry of *temp.*

The *k* loop is then placed inside the



*Sequence of nested toruses, analogous to the latitude lines on a sphere, projected from the hypersphere into three dimensions*



*Projected motion of toruses as the hypersphere is rotated, analogous to the projected motion of latitude lines during the rotation of a sphere*

intermediate loop that has index *j*. The *j* loop computes all four coordinates of the *i*th vertex of the rotated hypercube according to the procedure I have just outlined; in other words, it fills in all four entries in the *i*th row of *temp*. Finally the *j* loop is placed inside the outermost loop, which has the index *i*. The *i* loop calculates all 16 rows of *temp*, and when it is completed, *temp* gives the coordinates of all the vertexes of the rotated hypercube, resplendent in its new position. In order to display it on the computer monitor one more double loop is needed that replaces the old position coordinates of the hypercube in *cube* with the newly calculated coordinates from *temp*.

A hypercube has four dimensions but a display screen has only two. It is therefore convenient to stipulate that the first two dimensions, or coordinates, of the hypercube correspond to the screen coordinates. The simplest method for dealing with the third and fourth dimensions of the hypercube is to ignore them. The display technique I shall describe does just that, but it can be enhanced—and the resulting object can be projected in nearly demonic complexity—by making both the third and fourth dimensions somewhat more apparent.

To display a skeletal version of the hypercube, the program need only display its edges. Since the hypercube has 32 edges, the display section of HYPERCUBE need only draw the appropriate lines between 32 pairs of vertexes. But in what order? There are almost infinitely many possibilities, and so the answer is perhaps a matter of aesthetic and personal choice. Nevertheless, it is hard to resist drawing the edges as an Euler trail, after the mathematician Leonhard Euler. A pencil can trace such a trail on paper without being lifted from the paper and without tracing any line more than once. Consecutive edges of the hypercube drawn as an Euler trail have a common vertex.

Round, through, up and down races the Euler trail as it is drawn through the vertexes. Here is one that strikes me as quite pretty, given by the numbered vertexes of the hypercube connected in the following sequence: 0, 1, 3, 2, 6, 14, 10, 8, 9, 11, 3, 7, 15, 14, 12, 13, 9, 1, 5, 7, 6, 4, 12, 8, 0, 4, 5, 13, 15, 11, 10, 2, 0. These vertexes are stored in an array called *trail* with index *i;* the *i*th vertex in the sequence of 33 vertexes is designated *trail*(*i*). For each value of *i* there are instructions for looking up the first and second coordinates of both *trail*(*i*) and *trail*($i + 1$). The line-drawing command in one's programming language must then be invoked to connect the two points. The lookup and the line drawing are embedded in a single loop with index *i*, which draws a line from each vertex in the sequence to the next.

Now for the visual (and psychological) complications. There are two standard methods for presenting the third dimension of the hypercube. The orthographic method simply ignores the third dimension, and all the vertexes are projected directly onto the flat surface of the display screen no matter how far they are behind it. In one-point perspective the vertexes are projected onto the screen as though they were shadows cast by a point source of light centered on the screen and some distance behind the hypercube. Viewing the shadows on the screen is equivalent to viewing the hypercube from behind, but visually it is indistinguishable from a front view.

To achieve the effect of one-point perspective in HYPERCUBE one assumes that the third coordinate of a vertex is equal to the distance between the vertex and the display screen, in the direction of the imaginary point source of light. By solving for the sides of proportional triangles the program determines a multiplier needed to convert the first two coordinates of a vertex into screen coordinates. For example, if the imaginary light source is 20 units behind the screen, a vertex at $(5,-7,11,8)$ can be projected onto the screen by multiplying each of the first two coordinates by 20 and dividing each result by $20 - 11$, or 9.

I had dreaded including in the small space that remains a complete description of the process for creating stereoscopic images portraying the fourth dimension of the hypercube. There is a general technique for making stereoscopic images, and I hope to devote a future column to the subject. For the hypercube program, however, Banchoff and his colleagues have adopted a much simpler method. For each position of the hypercube make a new pair of images by applying *rot14* through an angle of three degrees in one direction and three degrees in the other. Dimension 1 is the direction parallel to the horizontal alignment of the viewer's eyes, and dimension 4 is the target of the exercise. The two small rotations nicely approximate the views of the hypercube from the eyes of the viewer: merely imagine the two lines of sight converging near the center of the hypercube at an angle of six degrees.

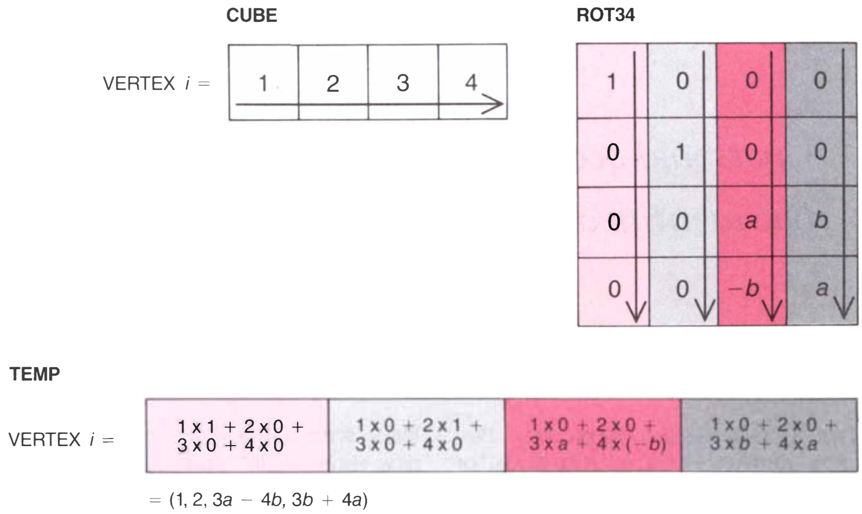
Readers who want to capture the thrills of 3D movies can make stereoscopic viewing glasses out of red and blue cellophane. In this case HYPERCUBE is run twice, once for each small rotation. The result of the first rotation is colored blue by the program and the result of the second is colored red. Readers need not be concerned about which is which if the eyeglasses are made to be invertible.

Personally I prefer not to struggle with cellophane, and I have learned to fuse stereoscopic pairs by shear force of will. The technique requires that the two rotated images be reduced in size and then translated to horizontally adjacent and nonoverlapping positions on the screen. They should be the same color, and so a monochrome screen is sufficient, and they should be no farther apart than the distance between the viewer's eyes. Do not stare at the images; look instead at some point between them and infinitely far beyond. The two hypercubes will appear to drift and jiggle toward each other like a pair of shy lovers until they fuse.

Even if the third and fourth dimensions get no special treatment, HYPER-

CUBE can generate images much like the ones shown in Banchoff's graphic sequence. With successive rotations through small angles in the third and fourth dimensions, readers may see the two crude toruses balloon, pinch off and regenerate much like their smoother cousins in the illustration of the rotating hypersphere.

The program HYPERCUBE had obviously caused the disappearance of my friend Magi. The happy ending to his four-dimensional dementia came with a telephone call. Not surprisingly, he spoke of wondrous things. "You probably think I'm crazy," he said. (The phrase is always a sure tip-off.) "I've just been floating around in the fourth dimension. I saw a cross section of my house sweep by. Then I moved in close and tickled my cat's kidneys..."




*How a vertex of the hypercube is rotated from the third dimension into the fourth*

I will spare the reader any further details of the conversation. Suffice it that I persuaded Magi to run HYPERCUBE no more and to keep further explorations entirely on the intellectual plane. He has followed my advice, he says, and now he professes to have made many marvelous discoveries through his artificially amplified insight. For example, he has come up with two posers that seem worth passing along.

Think for a moment about the following sequence of objects: a unit line, a unit square, a unit cube and so on. The $n$th member of the sequence is the $n$-dimensional analogue of the cube. Now try two mental experiments on the objects: draw the diagonal to the $n$-dimensional cube and inscribe an $n$-dimensional sphere within the $n$-dimensional cube. The diagonal stretches from one corner to the opposite one; what happens to its length as the number $n$ becomes progressively larger? What happens to the volume of the $n$-dimensional sphere, again as $n$ becomes progressively larger? Magi's answers seem hardly sane; I shall give them in next month's column.

In my January column I described two programs, CLUSTER and SUPERCLUSTER, that simulate the evolution of a star cluster. It heartens me to think that in at least a few thousand homes they have led to a new form of entertainment, temporarily edging out television. No doubt some of these armchair universes are unfolding as they should, but others may be developing problems. The fault is not in our stars but in ourselves.

Brian Davis of Ann Arbor, Mich., and Peter Fortescue of La Jolla, Calif., had trouble with the acceleration equations in SUPERCLUSTER. The difficulties are fixed, I believe, by replacing the force $f$ in the equation on page 13 with the acceleration $a$ due to the force. To get $a$ divide $f$ by the mass of the attracted star. Andrew M. Odlyzko of AT&T Bell Laboratories pointed out that the position coordinates in the table on page 15 are in multiples of 1,000 astronomical units (A.U.), not single A.U.'s. Our own universe will now unfold correctly.

When one views the live action of CLUSTER or SUPERCLUSTER on a display monitor, it is sometimes hard to tell which stars are in the foreground and which are farther back. Albert C. English of Delray Beach, Fla., and Peter Stearns of Lodi, Calif., have written special display programs that generate two images of clusters side by side, one as seen by the right eye and one as seen by the left. Readers able to manage the tricks of stereoscopic display will be able to view the clusters as they view the hypercube: in breathtaking depth.

Several readers had already written programs similar to SUPERCLUSTER, but they had applied the programs to our own solar system. The same application would also be feasible with SUPERCLUSTER. Those with the gumption can look up the mass, position and velocity of the 10 major bodies in the solar system for some reference time. One can then arrange to view the evolution of the entire system from above: wait a few minutes for the year 2000. Geoffrey L. Phillips of St. Louis, Mo., wrote a simulation for the earth-moon system that includes a small, massless space vehicle. Launching it from the earth in such a way that it begins to orbit the moon is no easy feat. Advanced practitioners might try launching a Voyager spacecraft on a grand tour of the gas giants that ends as it leaves the solar system.

William A. Hoff of Champaign, Ill., computed the time increment for the simulation dynamically by setting a variable called *dvmax* at the beginning of the program. In the course of the calculations of stellar motion the program always finds the maximum acceleration *amax* of a star. The next time increment is *dvmax* divided by *amax*. The technique prevents any velocity from exceeding *dvmax*.

In last month's column I gave an I.Q. minitest and posed several questions about numerical sequences. The first problem on the minitest is a good example of the ambiguity typically found in such problems. The problem was to complete the sequence 3, 7, 16, 35, .... Each term minus twice the preceding term gives the sequence 1, 2, 3, the second row of a pyramid. By this reasoning the missing term must be twice 35 plus 4, or 74. On the other hand, if a simple difference pyramid is constructed with three rows, the third row gives the sequence 5, 10, and it seems reasonable to complete the sequence with 15. The missing term must then be 69, but the programs described in the column would have missed this answer. The other answers to the test: $H$ is the missing letter; the missing word is "up"; the odd man out is "identity"; the unscrambled name of the town not in Italy is Madrid, and the correct visual analogy is number 2.

The two numerical sequences on page 12 are completed by 350 and 22 respectively. The first sequence on page 10 can be solved by applying a generalized difference rule, with $k$ equal to 3, and then a generalized quotient rule; the missing term is 324. The second sequence ought to defeat all but the most patient puzzle solvers who did not try to write SEQ. It can be solved by two quotient rules; the value of $k$ in the first rule is 5. The missing term is −65,551.

# BOOKS

## *A personal account of the epochal discovery that identified DNA as the genetic material*

by Salvador E. Luria

THE TRANSFORMING PRINCIPLE: DISCOVERING THAT GENES ARE MADE OF DNA, by Maclyn McCarty. W. W. Norton & Company, Inc. ($14.95).

Between 1604 and 1618 Johannes Kepler published the results of his lifelong studies of the motion of the planets, based on the remarkable naked-eye observations made by his teacher Tycho Brahe. Kepler's work established the heliocentric view of the planetary system and formulated its geometrical interpretation in the laws that go by his name: Kepler's three laws of planetary motion.

Thinking about Maclyn McCarty's *The Transforming Principle,* and reflecting on the history of the discovery it describes—the DNA nature of the gene-transforming substance of the pneumonia bacillus—I was struck by an analogy between Kepler's achievement and that of Oswald Avery, Colin MacLeod and Maclyn McCarty. Both advances provided factual and logically satisfying pieces of evidence that explained many converging but confusing observations. In both cases the factual conclusions were reached before they could be explained in terms of mechanisms.

More than half a century elapsed between Kepler's proclamation of the laws of planetary motion and Newton's interpretation of them in terms of forces acting at a distance and universal gravitation. Only nine years separated the classic 1944 article by Avery's group on the DNA nature of the transforming principle and the discovery of the double-helix structure of DNA by James Watson and Francis Crick. Science moves faster in the 20th century.

Newton's gravitational theory and Watson and Crick's double-helix model of DNA had an aura of abstract finality, of all-encompassing "revelation," that put them beyond the range of normal discovery—they were the leaps with which science breaks into new states of integration. In contrast, the achievements of Kepler and Avery were ordinary science—intelligent, stupendous, but not exceptional as feats of the human intellect—and neither of them took the scientific world by storm. Kepler's laws were probably unknown to Galileo in 1632; the discovery of the DNA nature of the extract that transformed the pneumonia bacillus got a mixed reception. It was disputed by biochemists, more or less ignored by bacteriologists, and was welcomed only by those geneticists who had been flirting with thoughts of genes and DNA. The background of this discovery, the history of its reception and the personalities of the participants make a fascinating story.

Maclyn McCarty's book is a genuinely intimate and knowledgeable account of the transforming principle's discovery and its bearing on the later deciphering of the DNA double helix. Of the previously published histories of the achievement it is complemented in these respects only by René Dubos' *The Professor, the Institute, and DNA.* The styles of these two authors are as distinct as one might find in the writing of two serious scientists. Dubos' prose is sensitive, vigorous and assertive. His narrative is marked by worship of Louis Pasteur and a devoted if puzzled admiration for Oswald Avery. McCarty's book is both a dry, low-key autobiography and a precise step-by-step report of the research that led to the discovery. It is a model of sober restraint, presenting the author and his colleagues as well as their critics without sentimentality. McCarty is almost touchingly precise in assigning to each statement and recollection the degree of certainty or uncertainty with which he can warrant it.

Out of the two books—Dubos' and now McCarty's—there emerge, painted with different palettes, the personalities in the story, particularly that of Avery, who was known as the Professor to most colleagues (an appellation that was shortened to Fess among his closest associates). The McCarty book also makes it clear that the TP discovery was outside the main lines of genetic and biochemical research, at the limit of the almost pathetically primitive (by 1986 standards) technology of the time. Simultaneously McCarty illustrates in a paradigmatic way the opportunistic process of research at the frontiers of science: the shrewdness required to grasp at disparate leads and approaches.

The story of the discovery can be summarized in a few paragraphs. It began in 1928 when the British bacteriologist Fred Griffith, puzzled by the frequent occurrence of multiple "types" of pneumococci—the pneumonia bacilli—in the lungs of a single patient, wondered whether the virulent types, each of them characterized by specific reactivity with an appropriate serum, could interconvert. Griffith injected mice with a mixture of nonvirulent bacilli together with heat-killed virulent bacilli of a specific "type." Some mice died, and in their lungs Griffith found live virulent bacilli of the type that had been killed. Many careful controls persuaded Griffith that the virulent pneumococci (referred to as S, for smooth colonies) had not just been revived: the nonvirulent ones (called R, for rough) had been "transformed" to virulence and had acquired the "type" of the dead ones.

Griffith's report fell on prepared soil at the Rockefeller Institute's pneumonia laboratory. Pneumonia, in those preantibiotic days, was the worst cause of death in the industrial world. Avery, the head of the laboratory, was a grand master of pneumonia research and one of the great men of bacteriology. Together with Alphonse Dochez, Avery first had isolated the substances that surround virulent pneumococci and are responsible for their virulence. Then he had shown that each type (I, II, III, ...) is associated with a specific substance and that each substance is responsible for the reaction of a type with its corresponding antiserum. This was a critical advance in pneumonia epidemiology, one that promised the possibility of specific vaccination.

Later Avery had brought Michael Heidelberger into the pneumococcus field and together they had proved that the specific capsular substances of pneumococci are complex sugars called polysaccharides. The discovery of the capsular substances had revolutionized immunology by proving that molecules other than proteins could elicit antibody responses in animals. At this point Avery had brought René Dubos to the Rockefeller Institute to search for enzymes that degraded the pneumococcal polysaccharides—and

---

*Philip Morrison's regular book reviews will appear again next month.*

The Library of Computer and Information Sciences

# Take any 3 books for only $1.00 each (values to $88.40)

(Publishers' Prices shown.) You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.


32486 $28.95 · 49353 $26.95 · 39678-2 $34.95 (Counts as 2 choices) · 39763 $26.95

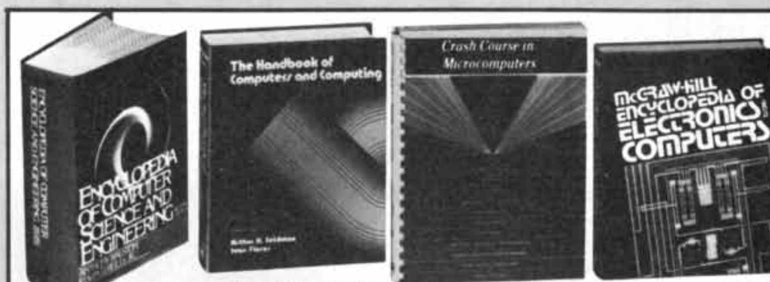
## Computer Science


39796 $14.95 · 52183-2 $55.95 (Counts as 2 choices) · 39880-2 $38.95 (Counts as 2 choices)

## Data Communications


42036 $29.95 · 65400 $26.95 · 74030 $22.95

## Systems Design

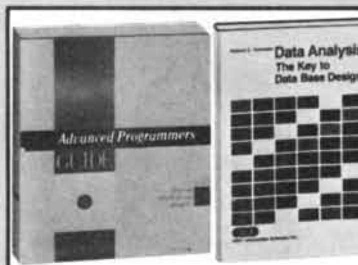



44904-3 $87.50 (Counts as 3 choices) · 52173-3 $77.50 (Counts as 3 choices) · 40862 $21.95 · 61613-3 $63.50 (Counts as 3 choices)

## Reference


85372-3 $77.90 (Counts as 3 choices) · 66008 $28.95 · 70723 $27.95 · 74781 $21.95

## Operating Systems




32272 $28.95 · 41620 $29.50

## Database


34150-2 $33.95 (Counts as 2 choices) · 38220 $24.95 · 55520 $19.95 · 68430 $19.95 · 32439 $15.95 · 39985-2 $39.00 (Counts as 2 choices) · 36455 $19.95

## Programming

**MEMBERSHIP BENEFITS** • In addition to getting 3 books for only $1.00 each when you join, you keep saving substantially on the books you buy. • Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices. • At 3–4 week intervals (16 times per year), you will receive the Library of Computer and Information Sciences News, describing the coming 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-DH4, Riverside, N.J. 08075** to obtain membership information and an application.

Scientific American 4/86


THE SKY WAS THE LIMIT.

AT&T has shattered the information barrier—with a beam of light.

Recently, AT&T Bell Laboratories set the world record for transmission capacity of a lightwave communications system—20 billion pulses of light per second. The equivalent of 300,000 conversations, sent 42 miles, on a hair-thin fiber of super-transparent glass. But that's really getting ahead of the story.

Actually, the 20-gigabit record is only one of a series of AT&T achievements in the technology of lightwave communications.

But what does that record mean?

### The Light Solution To A Heavy Problem

All of us face a major problem in this Information Age: too much data and too little information. The 20-gigabit lightwave record means AT&T is helping to solve the problem.

For data to become useful information, it must first be quickly, accurately and securely moved to a data transformer—a computer, for instance. Getting there, however, hasn't always been half the fun.

Metallic pathways have a limited transmission speed, sensitivity to electrical interference and potential for interception—factors that reduce the effectiveness of today's powerful computers. Factors that are eliminated by lightwave communications technology.

### Ten Goes Into One 20 Billion Times

Three primary components make up any lightwave communications system. On the transmitting end, a laser or light-emitting diode; on the receiving end, a highly sensitive photodetector; and in the middle, super-transparent glass fibers we call lightguides.

Installing these fibers is a major cost of a lightwave communications system. So, once installed they should stay put—increased capacity should come from fibers carrying more, rather than from more fibers.

Which brings us to the 20-billion bit-per-second story—about experimental technology that has the potential to upgrade installed fiber to meet any foreseeable capacity needs.

Using new, sophisticated lightwave system components, we multiplexed (combined) the outputs from 10 slightly different colored 2-billion bit-per-second laser beams into a single 20-billion bit-per-second data stream.

### Playing Both Ends Against The Middle

But, let's start at the beginning—the 10 distributed feedback laser transmitters.

These powerful semiconductor lasers can be grown to produce light of different, but very precise, wavelengths. The lasers we used transmitted in the 1.55 micron (infrared) range, with only minuscule fractions of a micron between their wavelengths. The purity and stability of the beams let us pack their ten colors into the most efficient transmitting region of our single-mode, silica-core fiber.

To make the original 10 beams into one, a fiber from each laser was fed into a new lightwave multiplexer—a prism-like grating that exactly aimed each beam into the single transmission fiber. Over 42 miles later, a second grating fanned the beam back into its original 10 colors for delivery to 10 exceptionally sensitive avalanche photodetectors—receivers that convert the light pulses back into electrical signals and amplify them many times.


20-gigabit multiplexer

A similar avalanche photodetector was the receiver when AT&T Bell Laboratories set the world record for unboosted lightwave transmission—125 miles at 420 million bits per second.

### From Sea To Shining Sea

System capacity is important. But system reliability is vital. Especially when the system is going under 10 thousand miles of water—and is expected to last for 25 years.

AT&T is going to build the first lightwave communications system under the Atlantic Ocean. A similar system is planned for the Pacific. In 1988, laser beams traveling through two pairs of glass fibers will carry the equivalent of 37,800 simultaneous conversations overseas, underwater, from the U.S. to Europe and the Far East.

AT&T has manufactured and installed lightwave systems—as large as the 780-mile Northeast Corridor and as small as single-office local area networks—containing enough fiber to stretch to the moon and back. And the capacity of each network is tailored to meet the unique needs of its users.

Systems being installed in 1985 will be able to grow from 6,000 up to 24,000 simultaneous conversations on a single pair of fibers.

AT&T is meeting today's needs with lightwave systems that are growable, flexible and ultra-reliable. And anticipating tomorrow's needs with a whole spectrum of leading-edge lightwave communications technologies.

**AT&T**

**The right choice.**

Dubos had succeeded: his enzyme specifically degraded type III polysaccharide and could protect mice from an injection of virulent bacilli.

When Avery read Griffith's report, he saw its importance and promptly got one of his colleagues, Martin Dawson, to repeat Griffith's experiments; Dawson's work confirmed Griffith's conclusions. Then Dawson and Richard Sia went further: they succeeded in proving that transformation of R pneumococci into specific S types could take place in a test tube rather than only in a living animal. This opened the way to precise experimentation. The next step, also taken in Avery's laboratory, was Lionel Alloway's finding that transformation did not require whole killed S bacterial cells: extracts of dissolved pneumococci, for example type III, could transform R cells derived from type I or II into live S cells of type III. The problem thereby became a biochemical one: the separation and identification of a "transforming principle" (TP) from the mixture of substances in the bacterial extract. Avery's name does not appear in any of the articles published by Dawson or Alloway, yet it is clear that Avery was and remained the driving force in the research.

Remarkably, no other investigative group jumped in to exploit Alloway's promising finding. At the Rockefeller Institute the TP research project moved at what by today's standards appears to have been a surprisingly leisurely pace, driven by Avery's conviction of the importance of identifying the transforming principle and constrained by the career requirements of a series of young medical researchers. Avery's laboratory was a most desirable center for the training in research of scientifically minded physicians, several of whom followed Alloway on the TP project. At times progress was also slowed by Avery's health problems, which led to a thyroidectomy for Graves' disease.

Enter finally Colin MacLeod. With him the effort to purify TP began in earnest, and yet it still progressed at a heartbreakingly slow pace. In their search for the transforming principle MacLeod and Avery methodically excluded one by one most components of their S-cell extract. Neither Dubos' enzyme, which broke down the polysaccharide, nor any of several enzymes that destroy proteins affected the TP activity. Tons of bacterial cultures had to be prepared for these tests and MacLeod invented the needed instrumentation as he went along.

By the time MacLeod left Avery's laboratory in 1941 many things had happened. For one thing, the arrival of sulfonamide drugs had made research on pneumococci less urgent if not altogether obsolete. He and Avery had to persuade themselves, as well as the institute's administration, of the desirability of continuing the study of TP. By then they had become convinced that TP must be nucleic acid—either RNA or DNA, substances then distinguishable only by uncertain chemical tests. MacLeod moved a mile and a half south from the Rockefeller Institute to the New York University Medical School. He remained in close contact with Avery but, surprisingly, did not pursue TP research in his own laboratory. It seems reasonable to assume that Avery was jealously keeping complete control of his pet project.

When McCarty replaced MacLeod, he may already have heard about the TP problem, but he was naive with respect to the substance of the work. His previous experience had been in pediatric bacteriology; still, he brought to the TP group the fresh drive and solid competence of a well-trained medical investigator. Methodically he and Avery refined the purification of TP, perfecting their techniques and pursuing every clue, such as the "stringy" appearance of the active fraction in an alcohol precipitate and the loss of activity on drying. Enzymes that degraded RNA did not affect this transforming activity. Most important, they proved that the transformed bacteria and their descendants became permanent sources of more TP of the same kind. The TP reproduced itself.

By 1943 McCarty and Avery felt certain that TP was DNA—in fact, pure DNA. It was not an easy certainty: in 1943 no one knew what pure DNA really was. Geneticists, particularly chromosome experts, had expected DNA to have something to do with genes, but they were outnumbered by chemists who claimed it could not. The chemists believed DNA was only a monotonous polymer, a molecule without importance, at most a dull portion of something that was called "nucleo-protein." And McCarty still had no reliable test for DNA. Altogether the proof of the DNA nature of TP was anything but "rigorous," the term by which Horace Judson described it in his summary of the TP story in *The Eighth Day of Creation*. Yet Avery and McCarty, joined by MacLeod, decided to go public.

No pressure or competition or sudden emergence of irrefutable data precipitated the decision to publish. In retrospect it appears that at this point Avery may have decided to exert the prerogative of the experienced leader. He exhibited the assertiveness that comes from the habit of success: a willingness to impose on a still confused mass of data a certainty that is emotional as well as rational. Such a source of certainty in science is unrecognized by those who believe certainty comes only after innumerable controls and attempts to disprove. The certainty Avery exhibited is more akin to illumination, a sudden vision projecting the possibility of an intellectual leap. What happened in the TP research at this point may have been the maturing of the conviction that the transformations observed in pneumococci were not just an idiosyncratic phenomenon of bacteriological interest but a stupendous lead to the chemical basis of hereditary specificity.

The article by Avery, MacLeod and McCarty, which appeared in the February 1, 1944, issue of the *Journal of Experimental Medicine,* is cautious in its conclusions but revealing in its opening sentence: "Biologists have long attempted by chemical means to induce in higher organisms predictable and specific changes which thereafter could be transmitted in series as hereditary characters." Note the precise wording, which could have satisfied any card-carrying geneticist.

Earlier in 1943 Avery had written his brother a letter, which fortunately became part of the public record because Avery showed it to Max Delbrück, who after Avery's death insisted that it be published. In the letter, reproduced in part in the McCarty book, Avery sounds cautiously but deeply stirred: "After innumerable transfers and without further addition of the inducing agent, the same active & specific transforming substance can be recovered far in excess of the amount originally used to induce the reaction. Sounds like a virus—may be a gene....the problem bristles with implications.... It touches genetics, enzyme chemistry, cell metabolism & carbohydrate synthesis, etc."

Whether or not Avery and his colleagues were aware of it, there was something in the air at the time those words were written. Just as the classical biochemistry of energy metabolism reached its apogee, marked by Fritz Lipmann's great review article of 1941, genetics was entering new and greener pastures. In December, 1940, George Beadle and Edward Tatum had reported their first results on the biochemical genetics of the bread mold *Neurospora crassa*—the work that was soon to lead them to put forward the one-gene, one-enzyme hypothesis. And in 1943 Delbrück and I had published in *Genetics* the first demonstration that bacterial variants originate by spontaneous mutation.

Bacteriologists were then by and

large oblivious of genetics. Many of them accepted uncritically hypotheses such as one put forward by the physical chemist Cyril Hinshelwood, which ascribed heredity and variations in bacteria to complex flow equilibriums in chemical reactions. In his 1976 book on Avery and the Rockefeller Institute, Dubos, one of the most biologically minded bacteriologists, remarks that as of 1945 he "could find in the literature only a few sketchy experiments to support the view that... some phenomena of bacterial variability nevertheless probably fall within the fold of classical genetics." It is a fact that most bacteriologists did not read *Genetics.*

Yet people such as Avery, MacLeod and McCarty certainly knew or sensed what was in the making. Theodosius Dobzhansky, in the second edition of *Genetics and the Origin of Species,* published in 1941, gave an account (the first one to come to my attention) of the transforming-principle story and its possible bearing on the study of gene mutations. In the summer of 1941, at Cold Spring Harbor, Dobzhansky developed a close intellectual relationship with Alfred Mirsky. Mirsky was a Rockefeller Institute biochemist who, in 1942, gave McCarty some useful leads on handling DNA and even collaborated with him in one set of experiments. Mirsky later became one of the severest critics of the Avery group and of their conclusion that TP consisted of DNA. He was critical partly on grounds of rigorous biochemical standards and partly, perhaps, because of a grudge over his abortive participation in the research.

Thus the Rockefeller Institute must have been abuzz with talk of biochemical genetics. It is reasonable to suppose (the supposition is not contradicted by anything in McCarty's book) that the profound biological significance of their TP work may, sometime between 1942 and 1943, have become so urgently clear to Avery, MacLeod and McCarty that the realization induced even the supercautious Avery to publish what was a *fact*—even though a fact validated as much by conviction as by hard evidence, a fact less fully convincing to experimentally minded critics than Kepler's laws might have been to Huygens and Newton.

Here I shall take a closer look at the authors of the 1943 article and its fateful statement. I have known all three of them personally. It is ironical to say that Avery stood out among the three. He was a "petite," gnomelike man with piercing but kindly eyes under an enormous forehead and a bald top, which he constantly stroked while he delivered to visitors his customary recitations of the pneumococcal transformation story (McCarty refers to these discourses as the "Red Seal Records"). His professional stature was apparent, as Dubos pointed out, in his complete lack of self-importance. He was unfailingly gentle, courteous in an unaffected way, yet prim and restrained. He had a sharp, teasing wit: I was told of his habit of greeting any colleague who had published a short preliminary note with the remark, "I have seen your advertisement." One day in early 1946 he walked into the Rockefeller Institute library before lunch just as I had found in the current issue of *Experientia* an article by the French bacteriologist André Boivin reporting DNA-mediated transformation in the bacterium *Escherichia coli.* When I showed Avery the article, he read it quickly, asked me to join him for lunch, and when a few of his co-workers sat down with us he whispered to them, "We seem to have continental support."

I visited Avery's laboratory at least three times between 1943 and 1945. I was interested in the connection between the TP work and my own research on bacterial mutations, wondering if the DNA might not transform a special class of mutant bacterial cells. (I worked for a few months on this topic and quit when later work in Avery's lab made my idea untenable.) Avery insisted that first of all he had to be absolutely sure of the chemical nature of the transforming principle.

So did McCarty, who by that time was searching for further proof of DNA involvement. McCarty's personality was radically different from Avery's. Intensely serious and alien to discourses, rather shy, laconic in explanation, he was a model of the serious medical researcher.

Jollier and more easygoing than McCarty was MacLeod, whom I came to know much better when, after the TP discovery, he became actively interested in bacterial genetics. In 1946 MacLeod asked Vernon Bryson and me (I was then spending a year at Cold Spring Harbor) to lead a genetics seminar at N.Y.U., where he headed the microbiology department. This turned out to be a delightful experience with a group of bright, intensely interested young people. In those few years bacteriologists had discovered genetics—thanks to the transforming principle. Dobzhansky was encouraging his best students to work in bacterial genetics (one of them, Harriett Taylor, joined Avery in 1945). And Tatum and Joshua Lederberg had discovered mating between *E. coli* cells.

The response of biochemists to the identification of TP with DNA was more cautious. Their bias was anchored in the belief that DNA, which when degraded yields just four components—the four "nucleotides" called *A, G, T* and *C* for short—had a simple repetitive structure (*AGTC-AGTC* or some other repeated sequence). Therefore, the chemists reasoned, DNA could not serve to encode any specific genetic information.

It is remarkable that this hypothesis, which had been proposed by P. A. Levene without any clear analytical evidence, had taken root in the mind of the few biochemists interested in nucleic acids—an unquestioned assumption that reminds me of the equally unchallenged assumption of astronomers before Kepler that planets should move in circles, whether around the earth or around the sun. The circle, according to Greek philosophers, was the ideal geometric figure. Levene's hypothesis, like that of Greek astronomers, appears to have been initiated by an unconscious desire to impose an ideal order of the simplest possible kind on another aspect of nature. The "sacred rage for order" (to use words from a poem by Wallace Stevens) when carried too far can generate hidden assumptions. The unmasking of such assumptions may in turn become the key to new advances.

Apart from Levene's specific model of DNA, a major obstacle to the acceptance of Avery's conclusion by biochemists, even after more solid evidence was provided by McCarty and by other Avery associates, was the lack of a plausible DNA "structure" on which to hang the specificity required of a genetic material. Proteins were "proteic," that is, multiform in structure. Their specificity was demonstrated, for example by serological tests. DNA instead remained mysteriously amorphous in the mind of biochemists. Bacteria, particularly pneumococci, might, it was suggested, be a freak class of organisms. Even if bacterial genes contained DNA, what did this prove about genes in fruit flies or human chromosomes?

The relatively long delay between Avery's discovery and its generalization to other genetic materials encouraged all kinds of questioning. It is still a tormenting recollection in my own scientific life that I (who had been so close to the TP story, had written a detailed report about it in 1947 and had even briefly worked on it) could in a 1951 paper suggest that the genetic material of a bacteriophage might be protein—a few months before Alfred Hershey and Martha Chase reported their brilliant experiments identifying DNA as the phage genome.

Geneticists had for years been told by some of the shrewdest cytologists, who were examining chromosomes by means of dyes more or less specific for protein or for nucleic acids, that DNA might be part of the genetic material. And yet others (besides myself) did not fully internalize the idea of the possible universality of Avery's discovery. Our hesitation is even more puzzling in retrospect when we consider that within months after the publication of the Luria-Delbrück test for mutations in *E. coli,* several genetic laboratories had taken up that lead. The failure of the Nobel committee to recognize the Avery work is particularly puzzling since one of the major experts in the chromosomal location and distribution of DNA was a member of the committee. Perhaps it was thought to be unseemly that the gene's identity should emerge from bacteriology.

The story of the fortunes of the TP discovery took a curious twist later when Gunther Stent, a molecular biologist and philosopher, suggested that Avery's discovery had failed to be appreciated because it was "premature." The doctrine of prematurity in discovery has provided a prime subject for arguments among historians of science in spite of the obvious circularity of the concept: is a discovery premature because it is not appreciated, or vice versa? More simply, does a discovery seem premature to those who missed its significance or failed to read or learn about it? The identification of TP as DNA was neither ignored nor unrecognized. It was questioned by some and embraced by others, and the universality of its impact remained uncertain. What Watson and Crick provided in 1953 was the chemical underpinning for the generalization.

Avery's work was unknown to Watson (he tells me) when, at the age of 18, he came to Indiana University as a graduate student. There he heard of pneumococcal transformation in my virology lectures and in Tracy Sonneborn's genetics course. Watson's fascination with DNA, which was both a rational and an emotional commitment (note the two key ingredients of science), dates from those days. It was this fascination that years later led him (with my encouragement and in spite of Delbrück's misgivings) to pursue the biochemistry of DNA. He ultimately stirred Crick to enthusiasm for that substance and drove the two of them to their great discovery, the DNA double helix: two complementary polynucleotide chains, each bearing complete directions for replication and each providing precise information for protein synthesis. DNA can be "transcribed" into RNA, which directs the synthesis of proteins.

This discovery stands at the center of biological science; it has both a towering finality and an aesthetic beauty that give it a stature comparable to the stature Newton's dynamics has enjoyed for the past three centuries in physics. Out of the double-helix model came a transformation of biology—the vision of a unity even more convincing than the great Darwinian synthesis. The clarifications of protein synthesis and of its regulation through control of gene action (a control itself mediated by specific protein-DNA interactions) are the high points of the new biology and now promise to solve the mystery of the development of complex organisms.

Reflecting on the prospects for further advance, one may be tempted to take an attitude of romantic pessimism: all that remains is either applications or epistemological disquisition. Such was the mood in physics around 1900—after Maxwell and Boltzmann, and just before Curie, Planck, Rutherford, Einstein and Bohr entered the picture. And so there is hope for the young biologists who dream of discovery.

Where might the "new" come from? I can think of two sources: areas that are today beyond the pale of molecular biology and areas where the current paradigm may again be found incomplete, as it was in physics. History, at least, offers the possibility that unwarranted hidden assumptions may be recognized in the present framework of molecular biology.

One area of persistent obscurity is the "cell theory." Can genes make a cell? Is the organization of a cell as a domain defined by a closed membrane fully coded in the genes, or does the membrane represent a self-priming pattern of organization? An answer to this question will require evidence for or against the possibility of synthesis of cells from subcellular elements. It may be sought by inquiring into the origin of polyphyletic cell types such as cyanobacteria or eukaryotic cells, which appear to have arisen from the merger of elements having different evolutionary origins.

The paradigm of molecular biology has already withstood several challenges that superficially seemed to bring into question its original formulation: the role of DNA as the genetic material and that of protein as the stuff of catalysis. Several viruses have RNA genomes, which are reproduced either through an intermediary DNA or directly as RNA. This, however, simply means that both kinds of nucleic acid can embody in their nucleotide sequences information capable of being copied. In fact, it now appears that a substantial portion of the DNA genome of animals probably originated with the copying of RNA into DNA sequences that were subsequently incorporated into the chromosomes. As for catalysis, the recent finding that an RNA molecule can in some instances catalyze its own restructuring does not shake the central notion that not genes themselves but only their products—mostly proteins but exceptionally also RNA—can function as catalysts for specific steps of cellular chemistry.

There is another twist beyond the simple picture of heredity based on mutation-prone but otherwise stable DNA sequences. It is the discovery that the response of the animal body to the innumerable variety of foreign substances involves a reprogramming mechanism for creating new DNA sequences—new genes—in the cells of the immune system. This discovery illustrates a subtle complexity of heredity based on DNA structure. The genetic endowment of an animal is more than a set of genes; it is a range of possibilities actuated during life by influences that selectively amplify clones of immune cells with the appropriately responding genes.

Molecular biology would be truly revolutionized by proof that substances other than nucleic acids can have a genetic role, that is, serve as molecular templates for the replication of themselves or for the production of nucleic acid chains. It has been suggested that amino acid sequences, primitive proteinlike molecules, might have had roles as templates in precellular stages of the evolution of life.

Whether a revolution in molecular biology is forthcoming, and whether any such revolution will emerge from novel findings at the molecular level or from resolution of epistemological quandaries, perhaps in dealing with the biology of the human mind, is for the future to tell. Avery and his coworkers were not after a revolution in science. They sought the answer to a specific question, the nature of TP, and it turned out to be a profoundly important answer. This is the normal path of the scientific enterprise: putting together the pieces of a puzzle whose design comes into existence only as the pattern grows.

Max Delbrück, speaking with Horace Judson, once pointed out that "science pretends that the scientist is immortal," a thought that caused him to worry lest it make scientists less human, less concerned about the mystery of human life. Is there for a biologist any true immortality other than having put into place, as Avery, MacLeod and McCarty did, a fragment that illuminates the overall pattern of the puzzle of life?

# "Not during a full moon I won't."

—Computer backup excuse #243

People can get very superstitious about when they do their computer backup. Especially if it gives them an excuse for not doing it at all.

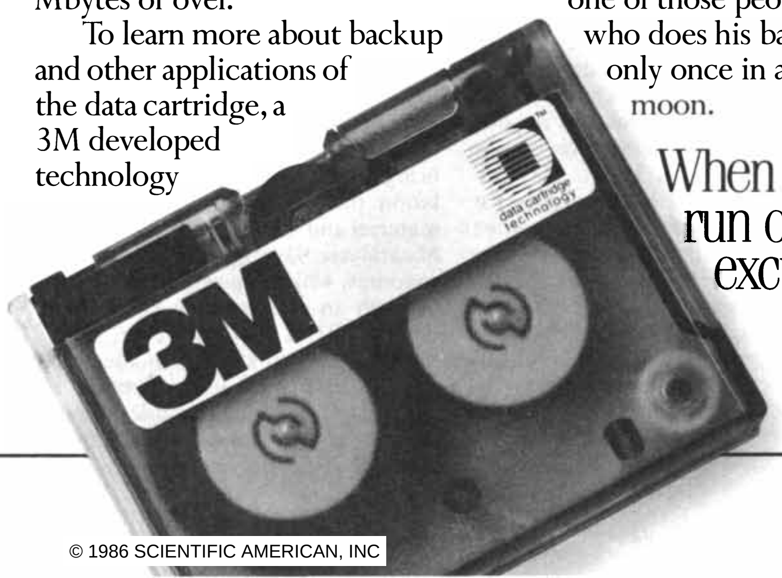
Because, after all, backing up is about as exciting as watching paint dry.

One way to take the curse off is to do it first thing in the morning. If you're one of those slow starters who has to have coffee and push some papers around for awhile when you get to work, that dead time could be perfect for backup.

As for how to do it, the floppy disk is fine if you have a limited amount of memory, and the data cartridge for 5 to 10 Mbytes or over.

To learn more about backup and other applications of the data cartridge, a 3M developed technology whose time has come, contact your local computer products dealer.

And maybe you won't be one of those people who does his backup only once in a blue moon.




When you run out of excuses.™

Established 1845 SCIENTIFIC AMERICAN April 1986 Volume 254 Number 4

# Can the U.S. Trust the U.S.S.R.?

*Neither country is beyond reproach in its observance of the SALT arms-control treaties. Soviet behavior can be understood in the broader context of superpower political and military competition*

by Miroslav Nincic

Almost from the moment the U.S. ratified the treaties drawn up at the first set of Strategic Arms Limitation Talks (SALT I), accusations of nonadherence have been made against the U.S.S.R. For instance, former Secretary of Defense Melvin R. Laird declared in 1977 that "the evidence is incontrovertible that the Soviet Union has repeatedly, flagrantly, and indeed contemptuously violated the treaties to which we have adhered." Most recently President Reagan, in 1984 and again in February and December of 1985, publicly accused the U.S.S.R. of violating a number of arms-control treaties, among them the SALT I Anti-Ballistic-Missile (ABM) Treaty and the SALT II Treaty.

Yet just as allegations of noncompliance have been reiterated from the time SALT I first entered into force in 1972, so does ambiguity recurrently afflict the evidence. Methods of monitoring compliance have always been imperfect; the information they provide is enveloped by a measure of uncertainty. Moreover, because each side is not willing to unduly limit its future options, simple blanket prohibitions are generally avoided in treaty language. Loosely worded or excessively qualified clauses, however, leave plenty of room for interpretation.

Even if definitive judgments regarding the U.S.S.R.'s fidelity to its promises can seldom be reached, a close look at the public record can nonetheless suggest some tentative conclusions. The record of the U.S. is also an instructive object of study. As arms-control delegations continue to meet at the negotiating table in Geneva and the American and the Soviet leaders prepare for their second summit meeting, it is appropriate to review the available evidence on Soviet compliance and to consider under what circumstances the Soviet Union engaged in dubious behavior, how far it went in each case and what determined the pattern of each side's compliance with its treaty obligations.

The most credible of the charges made by the Reagan Administration relates to the 1972 ABM Treaty. By the terms of the treaty, each side was allowed to deploy one ABM system at the national capital and one system at a launch area for intercontinental ballistic missiles (ICBM's). (In 1974 a protocol to the treaty further restricted each side to only one of the two originally acceptable areas.) The treaty also contained several provisions specifically designed to block each nation from acquiring a capacity to quickly deploy a more extensive ABM defense. One such provision stipulates that early-warning radars must be emplaced along the country's periphery and oriented outward. Such powerful radars, if they were built in the nation's interior, could otherwise cover a substantial area of the continental landmass and serve as battle-management radars for a nationwide ABM system.

In July, 1983, a Big Bird high-orbit satellite spotted a large new radar under construction in central Siberia. High-resolution imagery from a photoreconnaisance satellite subsequently confirmed that a phased-array radar was under construction near the village of Abalakova, north of the city of Krasnoyarsk. Unlike the other Soviet early-warning radars, which face outward and are situated on the nation's periphery, the radar near Krasnoyarsk is well inland (about 750 kilometers, or 465 miles, from the Mongolian border and some 1,600 kilometers, or 1,000 miles, from the Arctic Ocean) and is oriented so that it can cover a fan-shaped area opening toward the Bering Sea across several hundred thousand square miles of Russia.

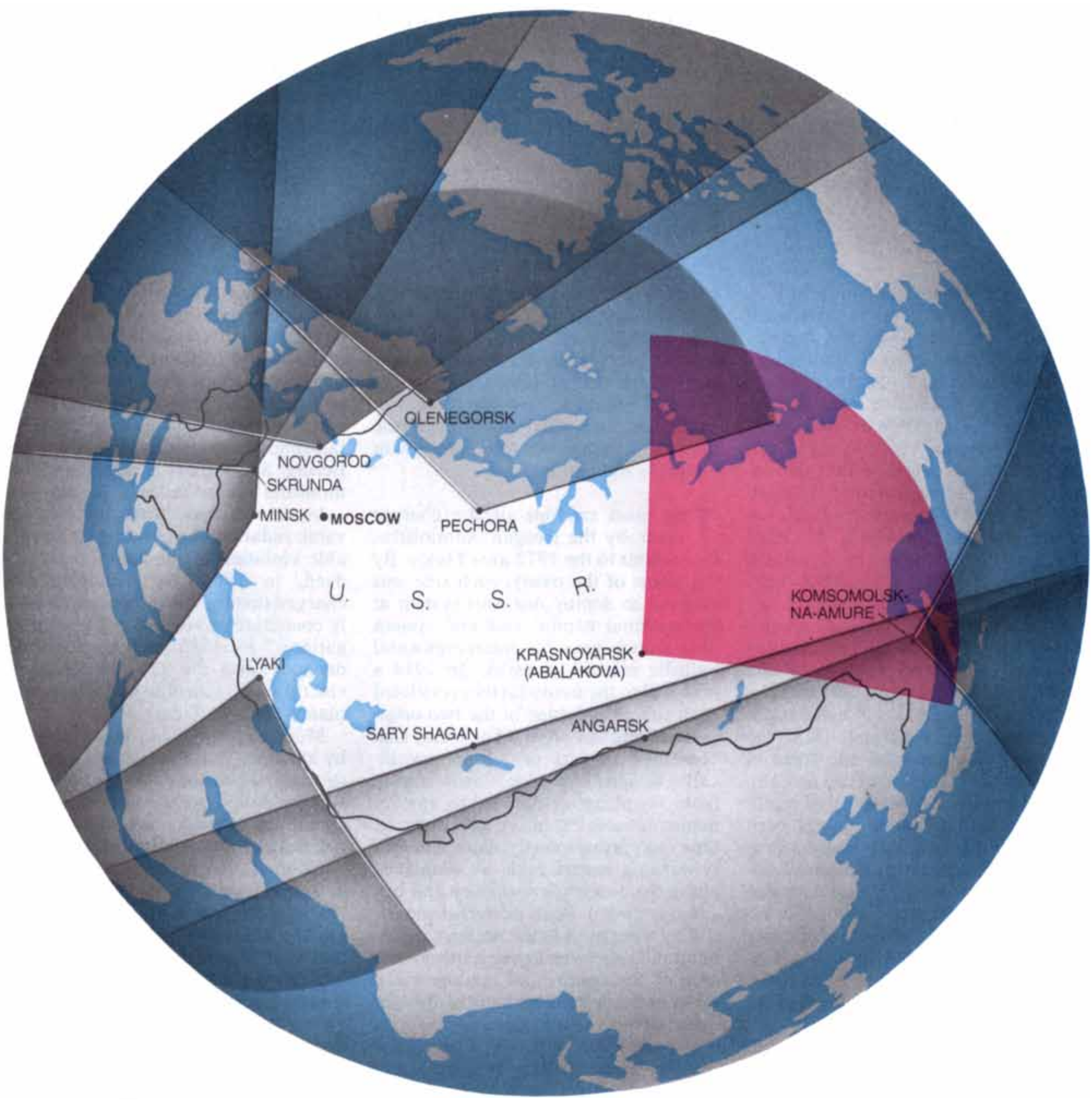
In view of these facts, the Krasnoyarsk radar can be considered a probable violation of the ABM Treaty. Indeed, in 1984 the Administration charged that the radar "almost certainly constitutes a violation of legal obligations." By 1985 all qualifiers were dropped and the Government succinctly stated that "the U.S.S.R. has violated the ABM Treaty."

Moscow responded to these charges by affirming that the radar is intended to track objects in outer space (a mission permitted by one of the agreed-on statements that accompanied the treaty), but the explanation fails to carry conviction. The radar is positioned in such a way that it would not make any useful contribution to tracking current space missions of the U.S.S.R. or for that matter of any other country. Its location would enable it to track only some 10 percent of earth-orbiting satellites—a paltry service for such an expensive construction project. Although the general opinion is that the Krasnoyarsk radar is in fact intended for the early warning of a ballistic-missile attack and not necessarily for an ABM defense, the Soviet contention that it is a legitimate system designed to track spacecraft is implausible.

This is not the first time the site of a Soviet radar has been considered to be contrary to the terms of the ABM Treaty, but in the other cases the infractions have been far less clear-cut. One earlier issue sprang from the discovery in 1975 that a new radar capable of ABM-related functions was being installed on the Kamchatka peninsula. The treaty permits the deployment of ABM subsystems if they are used for development and testing and as long as they are sited within "current or additionally agreed test ranges." The Soviets argued that the Kamchatka installation was on precisely such a range.

Prior to the conclusion of negotiations, however, the U.S. delegation had given the Soviet delegation a list of both sides' test ranges, and the list did not include the Kamchatka impact area. The Russians neither confirmed nor denied the list's accuracy or completeness, making the subsequent situation ambiguous. Because the location of the new ABM radar was not considered strategically significant, the U.S. relented and agreed that henceforth the Kamchatka radar would be considered part of an ICBM test range, as allowed under the treaty.

The Krasnoyarsk radar is not the only infringement of the ABM Treaty publicly alleged by the current Administration. In his report of February, 1985, President Reagan raised the




**SOVIET PHASED-ARRAY RADAR (*whose field of view is shown in red*) under construction at Abalakova, near the city of Krasnoyarsk, has been cited by the U.S. Government as an outright violation of the Anti-Ballistic-Missile (ABM) Treaty. The treaty, which bans nationwide ABM systems, requires that early-warning radars be emplaced on the periphery of the country and be outwardly oriented, as the U.S.S.R.'s other early-warning radars (*coverage shown in gray*) are. Such radars could otherwise keep watch over the nation's territory and serve as part of an extensive ABM system. The Soviet Union maintains that the Krasnoyarsk radar is intended not to provide early warning of a ballistic-missile attack but to track objects in space, a mission permitted by a mutually agreed-on statement appended to the treaty. Because of its position, however, the radar can track only a small percentage of earth-orbiting objects.**

matter of certain Soviet surface-to-air missile tests. The Soviets apparently have developed antiaircraft missiles, in particular the SA-12 hypersonic missile, which can also serve as an anti-tactical-ballistic-missile (ATBM) weapon. Although tactical weapons are not under the purview of the SALT agreements, ATBM systems could conceivably engage submarine-launched ballistic missiles (SLBM's), which, when they are launched into depressed trajectories, traverse flight paths similar to those taken by tactical missiles. Under certain circumstances ATBM systems might therefore have a limited ABM application.

More significant charges involved earlier antiaircraft systems. During 1973 and 1974 the U.S. observed that the radar associated with the Soviet SA-5 antiaircraft missile was being tried out in a new military role: to track strategic ballistic missiles in midflight—an important part of an ABM radar's function. The testing suggested an intention to make the system capable of engaging enemy ballistic missiles. This possibility had occurred to the U.S. during the negotiations, and an article of the ABM Treaty (later amplified by an American "unilateral statement") accordingly banned the testing of antiaircraft weapons in an "ABM mode."

The Carter Administration raised


**U.S. RADAR COVERAGE for the early warning of a submarine-launched-ballistic-missile (SLBM) attack will be provided by four PAVE PAWS installations; those at Beale Air Force Base and at Otis Air Force Base are currently operational and the other two are scheduled to become operational by 1987. The U.S.S.R. contends that the radars under construction at the Goodfellow and Robins bases contravene the ABM Treaty, which preserves deterrence by sustaining a mutual vulnerability to nuclear retaliation. Although the radars will radiate their signals outward, as stipulated in the treaty, the overlapping 240-degree fields of view (*red*) will blanket a significant part of the continental U.S. The Soviet Union asserts that such extensive overland coverage by advanced phased-array radars gives the U.S. the capability to support an ABM system that could defend American territory against a ballistic-missile attack.**

the matter with Soviet officials through diplomatic channels. Moscow denied any intention of altering the SA-5 for an ABM mission and pointed out that, even under the U.S.'s unilateral statement, such radars could be used for range-safety and instrumentation purposes. The U.S. recognized that the trials could have had this purpose and observed that, in any case, "much more testing, and testing significantly different in form, would be needed before the Soviets could achieve an ABM capability for the SA-5." The issue soon became an academic one since the Soviet Union thereafter halted the testing activities.

The current Administration has an additional qualm relating to Soviet observance of the ABM Treaty; it concerns a provision that prohibits the development, testing or deployment of mobile ABM systems. The Soviet Union is upgrading its single legal ABM system around Moscow to the so-called ABM-X-3 configuration. As a part of this improved system, the U.S.S.R. has developed a new set of modular radars, the "Flat Twin" radars. The units can be erected on a prepared site within months. The troubling implication is that this might represent a "mobile" ABM subsystem. Again, the situation is extremely ambiguous. Is a prefabricated house that can be rather quickly put together a mobile home?

Another set of charges leveled by the Administration primarily concern the second bilateral treaty on the limitation of strategic arms, SALT II. (The treaty has not been officially ratified by the U.S., but President Reagan has guaranteed that the U.S. will abide by it as long as the Soviet Union does.) The agreement, among other provisions, allows each side to test and deploy only one new type of ICBM. An appended "common understanding" specifies that existing missiles may be augmented by as much as 5 percent in length, diameter, launch weight or throw weight (payload). Any missile enlarged by a greater margin would be regarded as establishing a new class.

In October, 1982, at Plesetsk the U.S.S.R. began flight-testing a multiple-warhead, solid-fuel ICBM. The Soviets subsequently told the U.S. Government that the missile, designated the SS-24 by U.S. intelligence, was the new ICBM permitted under the treaty. In February, 1983, however, test flights of another missile, a three-stage, mobile ICBM, were also initiated at Plesetsk, placing the U.S.S.R. in apparent violation of SALT II. This missile, which has since begun to be deployed, is designated the SS-25.

The Soviets have denied that the SS-25 is a new type of missile. Rather, they maintain it is simply a modified version of an older ICBM, the SS-13, of which approximately 60 were deployed in the 1970's. The disagreement centers on measurement of the SS-13's throw weight. According to Moscow, the difference between the two missiles is not great enough to qualify the SS-25 as a new missile by SALT II criteria. In the opinion of a study prepared by the nonpartisan Arms Control Association, the data collected during the SS-13's flight-test program in the 1960's are inadequate to determine conclusively whether or not the difference in throw weight between the two missiles is more than 5 percent of the SS-13's throw weight.

Even the U.S. Government's 1984 report to Congress was comparatively cautious. It recognized that "the evidence is somewhat ambiguous," and declared deployment of the SS-25 to be only a "probable violation" of a "political commitment" to observe SALT II. Recently it was reported that the Soviet Union has carried out several test launches of the SS-13 to allow the U.S. to adjust its data, but it is not yet known whether the Administration feels reassured.

Washington states furthermore that the SS-25 represents a breach of SALT II in another way: by violating a clause that prohibits the testing or deployment of any single-warhead missile whose reentry vehicle is less than half the weight of the missile's "bus," the reentry vehicle along with the device from which the vehicle is released. The rationale behind this treaty provision was to make it difficult for either side to convert single-warhead missiles into multiple-warhead ones, since to do so would require development of relatively light reentry vehicles. In this way neither party could acquire a capacity for a rapid strategic "breakout" of warheads.

U.S. intelligence agencies seem to have concluded that in the case of the SS-25 the weight of the reentry vehicle falls somewhat short of the 50 percent minimum. The Soviets deny the claim; they imply that an instrumentation package was substituted for the reentry vehicle during testing, and that

| SOURCE OF CONTENTION | AMERICAN CHARGES | | SOVIET CHARGES | |
|---|---|---|---|---|
| | PAST | PRESENT | PAST | PRESENT |
| ICBM DEPLOYMENT | | SS-25 SS-16 | | MIDGETMAN (?) |
| ICBM THROW WEIGHT | SS-17 SS-19 | SS-25 | | |
| ABM CAPABILITY OF ANTIAIRCRAFT SYSTEM | SA-5 | SA-12 | | PATRIOT (?) |
| RADAR DEPLOYMENT | KAMCHATKA | KRASNOYARSK FLAT TWIN | SHEMYA ISLAND | PAVE PAWS |
| DELIBERATE IMPEDING OF VERIFICATION | COVERS OVER MOBILE ICBM's AND SUB FACILITIES | TELEMETRY ENCRYPTION SHELTERS OVER SS-25's | COVERS OVER MINUTEMAN AND TITAN | |
| ABM-SYSTEM-COMPONENT TESTING | | | | SDI-RELATED TESTS(?) |
| CIRCUMVENTION THROUGH OTHER STATES | | | | NATO DEPLOYMENT OF PERSHING II's AND CRUISE MISSLES |

**ALLEGATIONS of noncompliance with arms-control agreements display a certain consistency, with charges made in the past few years relating to dubious actions that in many cases are comparable to actions that drew similar charges a decade or so ago. The background color indicates the particular treaty to which the objections listed here refer: the Interim Agreement (*light color*) and the ABM Treaty (*dark color*), both drawn up in 1972 at the first set of Strategic Arms Limitation Talks (SALT I), and the SALT II Treaty (*gray*), signed in 1979 but never officially ratified by the U.S. Question marks designate instances in which the U.S.S.R. could justifiably charge the U.S. with infringement of treaty provisions unless the treaties are amended or the U.S. changes its plans. In the case of the Patriot antiaircraft system the purportedly illegal activity has apparently been discontinued.**

the package is lighter than the vehicle would be.

To put the controversy in perspective, one should remember that claims of Soviet noncompliance with size and weight restrictions have been made before. Of the 1,409 launch-ready ICBM's allowed the U.S.S.R. under SALT I, the U.S. considered the 1,096 SS-11's to be "light" ICBM's, while 313 SS-9's were classified as "heavy." Because the U.S. had conceded the Soviet Union more launchers in order to counterbalance the U.S. lead in total number of warheads, it seemed important not to let the U.S.S.R. disrupt the negotiated rough strategic equivalence by increasing the size and hence the warhead-carrying capacity of their weapons.

The Soviet SALT delegation, however, resisted any discussion of missile size and weight, and it was left up to the American delegation to proclaim, through a unilateral statement, that an ICBM would be considered heavy if it had a volume "significantly greater than that of the largest light ICBM now operational." The American understanding was that heavy ICBM's were not to be substituted, in any significant measure, for those in the light class. In spite of the enunciated U.S. position, the Soviet Union undertook to replace a number of SS-11's with larger SS-17's and even larger SS-19's. This resulted in an increase in missile throw weight of 186 percent in the first case and more than 300 percent in the second (although in both cases the throw weight remained substantially below that of the heavy SS-9).

Was this a treaty violation? Technically it probably was not since the definition of a heavy ICBM and the obligation not to build one were rooted in a unilateral U.S. interpretation of the treaty. Still, political predilections guided opinions in opposite directions. Senator Jake Garn felt that "the deployment of the SS-19 would have justified, indeed, should have demanded U.S. abrogation of the accords." Yet then Secretary of State Henry Kissinger rejoined that "it is at least open to question whether the United States can hold the Soviet Union responsible for [the U.S.'s] statements when the Soviet Union has asserted that it does not accept this interpretation."

Another, more recent problem involving the SS-16 mobile ICBM also stemmed from an unclarified definition. Under the terms of the SALT II Treaty the SS-16, originally developed in the 1970's, could no longer be produced, tested or deployed. Yet the issue of what was to be done with those SS-16 missiles already in existence (estimated to number between 30 and 80) was left unresolved. Although they were clearly not to be deployed, the exact meaning of "deployed" was never specified. The U.S.S.R. presumed it had fulfilled its obligation by putting away the existing SS-16's at the Plesetsk missile-testing grounds. Is a missile deployed if it is mothballed at a test range? The Reagan Administration argued that it was. In any event, the missiles have since been removed and put into storage, and so at least one source of controversy has been eliminated for the Administration.

**SOVIET SA-5 MISSILE was designed to intercept high-altitude enemy aircraft. During 1973 and 1974, however, the U.S. observed the radar associated with the SA-5 tracking ballistic missiles, a function of ABM radars. The ABM Treaty forbids the testing of antiaircraft systems in an "ABM mode," and accordingly the U.S. Government objected through appropriate diplomatic channels. The Soviet Union denied any intent to endow the SA-5 system with an ABM capability, but soon thereafter it ceased the controversial activity.**

Given the wariness that the two signatories to the SALT agreements have toward each other, it is not surprising that each side continuously monitors the compliance of the other through intelligence-gathering devices such as sophisticated photoreconnaissance cameras, radars and eavesdropping radio antennas based in space, on land and at sea. These devices are collectively known as national technical means of verification [see "The Verification of Compliance with Arms-Control Agreements," by David Hafemeister, Joseph J. Romm and Kosta Tsipis; SCIENTIFIC AMERICAN, March, 1985]. In order to facilitate the operation of these nonintrusive methods of monitoring treaty compliance, both parties pledged to limit the extent of concealment with which they surrounded their strategic weapons. An important provision of SALT II in this regard limited the encryption of telemetry.

Telemetric information is test-performance data that are transmitted as a stream of radio signals from the missile to monitoring stations while the missile is in flight. Because each side routinely intercepts the other's signals to gather military intelligence, each side encrypts part of its telemetry to deny the other secret information.

A common understanding appended to the SALT II Treaty states: "Each Party is free to use various methods of transmitting telemetric information during testing, including encryption, except that... neither Party shall engage in deliberate denial of telemetric information, such as through the use of telemetry encryption, whenever such denial impedes verification of compliance with the provisions of the Treaty." The statement implies that information bearing on treaty compliance can be distinguished, at the practical level, from information that does not (although it may still be sought for intelligence reasons), and that different transmission channels or modes can be assigned to the two types of data.

By most reports about 60 percent of Soviet telemetry is regularly encrypted, but the figure has been significantly greater for tests of the most recent ballistic missiles: the SS-24 and the SS-25 ICBM's, as well as the SS-N-20 SLBM. According to the Reagan Administration, the increase in Soviet encryption constitutes "an example of deliberate impeding of verification of compliance."

There is no doubt that the greater

the level of encryption, the stronger the suspicion that the Soviet Union is undermining SALT II. Yet, as was explained to the Senate by General Charles A. Gabriel, the U.S. Air Force Chief of Staff, "it is misleading to use the level of encryption as a guide for determining how well we can verify what they are doing. More central is the nature of the encryption, that is, what missile-test functions are being encrypted and what functions are being transmitted in the clear." In other words, as long as the signals being encrypted are not those that describe missile characteristics regulated by treaty provisions, the U.S.S.R. is not impeding SALT II verification.

As in the case of most charges regarding the second of the SALT agreements, the problem can be traced back to SALT I. Accusations of Soviet telemetry denial were made in the 1970's, but the legal basis for the complaint was weaker then. Although SALT I prohibited deliberate concealment and interference with national technical means of verification if they impeded compliance monitoring, no explicit mention was made of telemetry.

Charges of deliberate impeding of verification have also been made with respect to shelters that have been erected over deployed strategic weaponry, namely SS-25's. Here again, this Soviet practice had been critized before. A vague clause is at the root of the misunderstanding. The SALT treaties permit "current construction, assembly, conversion, or overhaul practices" to continue in spite of the restriction against compliance-hindering measures. The explicit exemption allowed the Soviets to argue that their strategic submarine construction and refitting facilities could remain covered because they had been covered before the SALT I Treaty was signed. As for assertions that they also covered land-mobile ICBM launchers in contravention of SALT I, the Soviets pointed out that the agreement (apart from an American unilateral statement) did not mention this type of weapon system and so they felt no obligation to act as if it had. They also recriminated that the U.S. had erected shelters over its own ICBM launchers. Nevertheless, the U.S. officially filed a complaint on both counts and the Department of State later reported, somewhat enigmatically, that "there no longer appears to be an expanding pattern of concealment activities associated with strategic weapons programs."

The review of present and past allegations of misconduct makes it clear that incontrovertible, provable breaches of treaty provisions are not to be expected from the U.S.S.R. From the Soviet perspective it is much easier not to sign an agreement than it is to incur the consequences of manifestly violating its provisions. Nevertheless, it does appear that the U.S.S.R. maneuvers perilously near the brink of outright noncompliance, without (except perhaps in the Krasnoyarsk case) going demonstrably beyond the brink. This kind of behavior can hobble progress on arms limitation, but it is precisely the conduct that should realistically be expected. The costs and risks of outright cheating are usually excessive, yet strategic imperatives are too great not to attempt to tread the extreme limits of the permissible.

From this point of view the relevant question is not whether the Soviet Union engages in dubious tactics but how close to the verge of unambiguous noncompliance it will go and what determines the degree of boldness in this regard. Some general principles and tendencies seem fairly apparent.

The U.S.S.R. can strain the integrity of an arms-control agreement in two ways. To begin with, it can engage in an illegal activity that the U.S. cannot positively identify as such; available evidence might suggest a violation without definitively proving it. In this case uncertainty surrounds actual Soviet behavior rather than the language of the agreement. The other option is to interpret a treaty provision in a way that is both self-serving and contrary to U.S. interests, but without violating the letter of the agreement. In either case Soviet behavior is obviously prompted by underlying political or military objectives.

A large part of the Soviet Union's compliance brinkmanship may in fact be perpetrated to gauge the state of U.S. monitoring capabilities and to test the limits of American forbearance. Much superpower interaction is in essence a probing game designed to extract intelligence or to calibrate behavior according to expected response. Several of the apparent SALT violations are explicable in these terms. With regard to SALT I, for example, the use of the SA-5 radar to track missiles in flight may have been intended to find out how the U.S. defines "testing in an

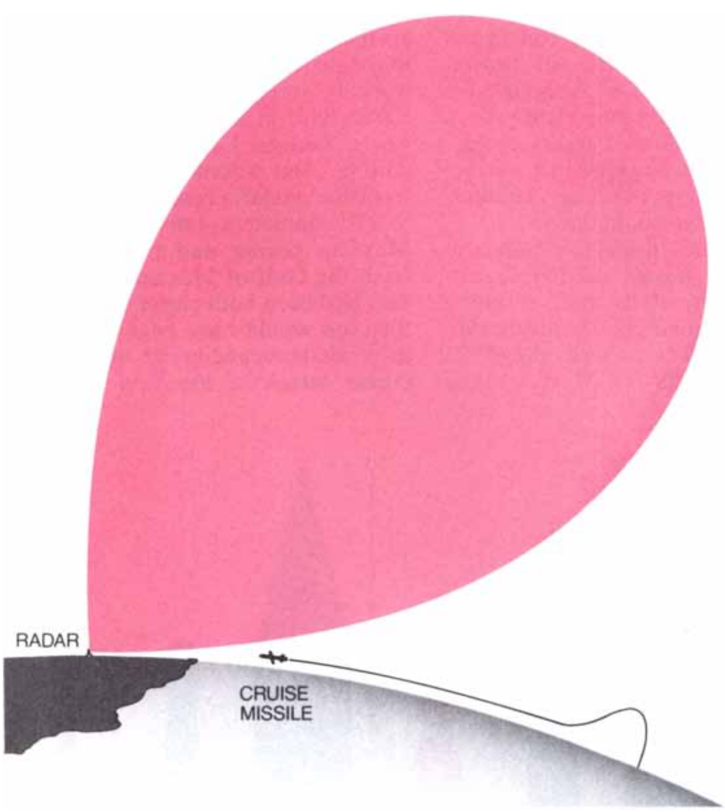
**U.S. NAVY SHIP *Observation Island* has a COBRA JUDY phased-array radar mounted on the flat deck near its stern. The sophisticated radar collects intelligence data on flight tests of Soviet missiles. It, along with other devices such as photoreconnaissance satellites and radio antennas, is a "national technical means" by which the U.S. can monitor Soviet compliance with SALT treaty obligations.**

ABM mode." Similarly, the extensive encryption of telemetry for the Soviet Union's fifth-generation ICBM's could have had as one of its purposes the assessment of the telemonitoring and analysis capabilities of U.S. intelligence. Indeed, exploratory probing of this kind is an important facet of superpower competition and is carried out continually in other contexts, such as when one side deliberately provokes a response from the other by having military aircraft loiter just outside the other's airspace.

More generally, however, Soviet activity at the treaties' margins has the straightforward objective of either obtaining a military advantage or of offsetting one achieved by the U.S. The Krasnoyarsk radar is a case in point, and it is particularly significant in that it is the most convincing instance of a direct breach of SALT obligations.

Early-warning radars are generally a response to a perceived offensive-missile threat. The Krasnoyarsk radar may therefore have been built to redress a strategic balance that had threatened to tip in the U.S.'s favor as Trident submarines are deployed in the North Pacific. Because the radar is oriented toward the northeast, it would be able to detect Trident SLBM launches from the Bering Sea or the Gulf of Alaska; it thereby closes a vulnerable gap in Soviet early-warning-radar coverage. Even the U.S. Central Intelligence Agency has apparently recognized that the contentious Soviet radar is not well suited for ABM battle management and is most probably intended to serve as an early-warning sensor of SLBM attacks.

Nevertheless, because of its location it probably violates the ABM Treaty; indeed, its location presents an enigma. In principle the radar could have been built closer to the northeastern periphery, say near the Sea of Okhotsk. An early-warning radar on the seaboard would have met the conditions of the treaty. Moreover, it would have been able to detect incoming ballistic missiles sooner than a radar far from the coast. Why then was the radar emplaced near Krasnoyarsk?




**CRUISE MISSILE can elude prompt detection by flying below radar coverage. An early-warning radar emplaced near the coastline is therefore much more vulnerable to a sea-launched cruise-missile attack than it would be if it were inland, where the missile would have to penetrate several layers of the nation's air-defense system in order to reach the radar. This military reason for placing an early-warning radar in the interior of the country may explain why the U.S.S.R. is constructing such a radar near Krasnoyarsk in central Siberia in spite of vehement U.S. accusations that it is thereby violating the ABM Treaty.**

The Soviets have cited environmental necessity, and specifically the absence of permafrost, to explain the radar's location. But this is not very convincing: the Soviet Union does not shy away from rough weather when military interests are at stake. Another explanation is that Krasnoyarsk is conveniently situated for major construction work: it is not far from the city of Novosibirsk and is connected to the nation's transportation network through a spur of the Trans-Siberian railway. Still, it is unlikely that mere economic convenience would bring the Soviets to expose themselves to credible charges of cheating. The radar's location, like its probable military role, must also be understood in its proper strategic framework.

One explanation might be found in the radar's ability to support a Soviet launch-on-warning policy. Its proximity to missile launchers (it is within a few hundred miles of three of the U.S.S.R.'s six SS-18 ICBM fields and near an SS-11 missile field and an SS-20 missile field as well) would enable it to give an unequivocal indication that an enemy missile attack was aimed at the nearby missile fields. It could thereby provide a reliable basis for a decision to launch the SS-18's (and perhaps the SS-20's) before they were hit. If that is the case, the Soviet launch-on-warning strategy may have stemmed from the reduced warning time associated with the SLBM's that can be fired from Trident submarines.

Perhaps an even stronger reason for the radar's inland site is the growing threat of U.S. submarine-launched cruise missiles (SLCM's). From the Soviet point of view the most direct danger would come from the Tomahawk land-attack SLCM, which can be equipped with a nuclear warhead or with a 1,000-pound high-explosive, conventional warhead. Because cruise missiles can evade air-defense radar by flying at low altitudes, they could be used to destroy early-warning systems in a surprise attack and thereby open a "window" through which an SLBM assault could then be launched. If, moreover, this were done with the nonnuclear Tomahawk (and there would be no reason to use a nuclear warhead against a relatively small, vulnerable and exposed target such as an early-warning radar), the Soviet military position would be significantly weakened at the very start of a military confrontation without transgression of the nuclear threshold. Such a move might give the U.S. increased leverage during a particularly volatile crisis. By establishing the radar near Krasnoyarsk, Moscow may have wanted to deny the U.S. such a strategic option.

None of this makes the radar's location any more legitimate under the ABM Treaty. The point is merely that such developments can only be understood when they are considered in the complete matrix of Soviet-American competition. Although the U.S. has adhered more rigidly to a strict interpretation of treaty obligations, it has also engaged in questionable conduct of its own when strategic imperatives have so dictated.

For example, projected improvements in the Soviet submarine-missile force (particularly those resulting from the Typhoon-class strategic submarine, the SS-N-20 and SS-NX-23 SLBM's and the SS-NX-21 SLCM) may account for the Government's decision to construct two new PAVE PAWS early-warning radars at Goodfellow Air Force Base in Texas and Robins Air Force Base in Georgia. The former will be 260 kilometers (160 miles) from the Atlantic coast, the latter will be 220 kilometers (140 miles) from the Mexican border and twice that far from the Gulf of Mexico. If these radars had been built closer to the coast, they too would have been vulnerable to a sea-launched-cruise-missile precursor attack of the kind that may have motivated construction of the radar near Krasnoyarsk.

Both new American radar installations could be considered in contravention of the ABM Treaty not only because of their "inland" locations but also because their combined field of view could cover as much as two-thirds of the continental U.S. On these grounds the U.S.S.R. has officially charged the U.S. with a violation of the treaty.

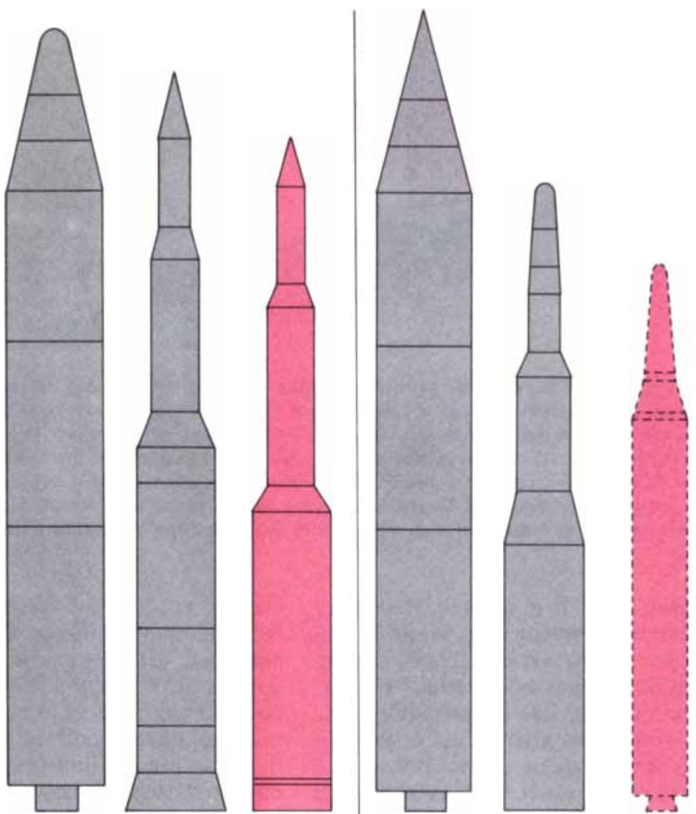
This is not the only instance where the Soviet Union has pointed an accusing finger at the U.S. In 1975 the Soviets objected that a U.S. radar on Shemya Island in the Aleutian chain might be a prohibited ABM radar. Washington's response, that the radar was used for space tracking and early warning, was apparently accepted by Moscow.

The U.S. has also risked exposing itself to charges of upgrading antiaircraft systems for a potential ABM role. In April, 1984, the *New York Times* reported that the Patriot missile system (a frontline antiaircraft weapon system) was being endowed with the capacity to shoot down Soviet missiles. While this may have been a response to the Soviet SS-20 deployments in western Russia, the practical difference between a capacity to destroy intermediate-range ballistic missiles and ICBM's is blurry.

Less ambiguous is the situation that would be created with regard to SALT II (assuming a desire by both nations to continue observance) if the U.S. goes ahead with the single-warhead, Midgetman ICBM program. In that event the U.S. would be contravening the stipulation permitting only one new ICBM: that option was expended with the deployment of the MX (Peacekeeper) ICBM. The most flagrant disregard of a treaty currently in force would occur once operational testing of the space-based systems implied by President Reagan's Strategic Defense Initiative (SDI) begins. Under the terms of the ABM Treaty each party is pledged not to develop, test or deploy ABM systems or components that are space-based. Laboratory research on such systems is permissible, but once actual testing begins the ABM Treaty would be, by any reasonable interpretation, violated.

The depth of the rivalry is such that neither superpower is prepared to sacrifice major strategic objectives for a strict adherence to arms-control treaties. Because of their priorities, it is understandable that both sides frequently skirt the edges of the permissible.

One final consideration is that issues of compliance and noncompliance are very much colored by U.S.



| | SS-24 | SS-13 | SS-25 | MX | MINUTEMAN II | MIDGETMAN |
|---|---|---|---|---|---|---|
| YEAR FIRST DEPLOYED | 1986 (?) | 1972 | 1985 | 1986 | 1965 | EARLY 1990's |
| MAXIMUM RANGE (KILOMETERS) | 10,000 | 9,400 | 10,500 | 14,000 | 12,500 | ≈11,000 |
| NUMBER OF WARHEADS | 10 | 1 | 1 | 10 | 1 | 1 |

**CONTROVERSIAL "NEW" MISSILES (*color*) have drawn charges of noncompliance from both sides. The SALT II Treaty permits flight testing and deployment of only one new intercontinental ballistic missile (ICBM). Modifications to an existing ICBM that stay within 5 percent of its original length, diameter, launch weight and throw weight (payload) are not regarded as constituting the development of a new ICBM. Both sides are taking advantage of the option to deploy a single new type of ICBM: the U.S.S.R. will deploy the SS-24 and the U.S. will deploy the MX (Peacekeeper). The Soviet Union, however, has recently begun to deploy another ICBM, the SS-25. The U.S. maintains that this contravenes the SALT II provisions. The U.S.S.R. denies the charge on the ground that the SS-25 is actually a permitted modernization of an earlier ICBM, the SS-13; it countercharges that the U.S., in fact, has transgressed the treaty by undertaking the development of the Midgetman ICBM. It is not known whether the U.S. Government will treat the Midgetman as a modified version of the Minuteman II ICBM in an effort to remain within the bounds of the SALT II Treaty.**

politics. A changing alignment of international political forces or shifts in domestic political moods will make what was once considered tolerable appear unacceptable. For example, although the same kind of compliance issues that currently preoccupy the Reagan Administration have been brought up repeatedly since the first SALT treaty was signed, both the Ford and the Carter administrations were satisfied with the Soviet record of compliance.

In 1975 President Ford explained that he had "investigated" the allegations and concluded that the Soviets "have not violated the SALT agreements. They have not used any loopholes." Somewhat more obliquely, the State Department maintained in 1978 that, although the U.S. had raised questions regarding several Soviet activities, "in each case, the activities in question ceased or additional information has allayed the concern."

More to the point, U.S. denunciations of the U.S.S.R. (and vice versa) are part of the diplomacy of strategic competition. For example, many observers have conjectured that the Administration's charges against the Soviet Union are designed to provide it with some immunity from accusations of its own potential treaty violations when components of the SDI program enter the stage of operational testing, or if the Midgetman missile comes to be deployed. Similarly, the U.S.S.R. may have hoped to besmirch the arms-control record of the U.S. by presenting the deployment of intermediate-range, nuclear-capable Pershing II and cruise missiles in Europe as a violation of the SALT II Treaty, in spite of the fact that the treaty deals exclusively with weapon systems of intercontinental range. The Soviets based their argument on an article in the treaty that prohibits circumvention of the treaty's provisions "through any other state or states," presumably allies. By deploying intermediate-range nuclear weapons that can strike targets in western Russia from bases in countries of the North Atlantic Treaty Organization, the U.S., the Soviets maintained, has effectively gone around the agreed-on numerical limits placed on strategic weaponry.

The diplomatic leverage that the formulation and timing of public charges of superpower misconduct can bring to bear is most recently illustrated by a report that the Soviets were told in April of last year they might be taken "off the hook" on the Krasnoyarsk violation if they consented, as a quid pro quo, to a redefinition of ABM Treaty provisions that would allow some SDI-related tests to proceed beyond the laboratory stage. The Soviets in turn suggested the Krasnoyarsk radar might be scrapped if the U.S. abandoned modernization programs for its early-warning radars in Fylingdales Moor in England and Thule in Greenland—an offer the U.S. rejected.

Because the U.S.S.R. (and the U.S.) will often hover near the edge of noncompliance with arms-control treaties, should one conclude that such treaties are in the final analysis worthless? Not at all. The SALT agreements do impose constraints on military growth—loose constraints, but constraints nonetheless. From the time SALT II was signed, the Soviet ICBM inventory has remained nearly constant. To allow for the deployment of new weapons under SALT ceilings, between 1972 and 1985 the Soviets withdrew 1,007 land-based and 233 submarine-based ballistic missiles from their active missile inventory and dismantled 13 ballistic-missile submarines. Without these limits they could simply have allowed the weapons to accumulate, as they had done in the past. They have also maintained a maximum of 10 warheads per ICBM, which is considerably less than the number they could theoretically deploy. (The SS-18, for example, can accommodate between 30 and 40 reentry vehicles.) Indeed, as was stressed to the Senate Armed Services Committee by Lt. Gen. John T. Chain, Jr., former director of the State Department's bureau of politico-military affairs, the U.S.S.R. might be in a more favorable position than the U.S. to move ahead in a renewed arms race if the treaty limits were suddenly declared void.

Because arms-control agreements do make a difference, they should be pursued. Yet in negotiating them care must be taken to guard against self-serving interpretations as well as the kind of probing activity that undermines their effectiveness and whatever political support they enjoy. How can this be done?

The obvious solution is to call for more precise specification of treaty obligations and for improved means of verification. Certainly both are needed to reduce the leeway for undesirable behavior. Loopholes must not be left open; for example, the matter of when telemetry encryption impedes verification should be clarified. In addition one should try to go beyond current national technical means of verification; Moscow's growing willingness to accept a measure of intrusive verification, such as on-site inspections, is encouraging in this respect. Still, there are inherent limitations to both specificity and verifiability; in practice the problematic issues often concern the most appropriate way of dealing with objectionable but not demonstrably illegal activities.

It has become quite clear that little is to be gained by the shrill incriminations and recriminations that are often the preferred approach to East-West relations. This may provide some symbolic gratification, but histrionics and bluster simply do not alter Soviet behavior. Soviet compliance appears to have been no worse at the time of detente than it was during the hard-line, confrontational period of President Reagan's first term. The U.S. should certainly try to thwart Soviet actions that deviate from a reasonable view of what existing agreements allow, but just as the proper tone is not one of self-righteous indignation, so the proper forum for discussion is not, at least initially, the public arena.

The appropriate setting for compliance diplomacy is the Standing Consultative Commission (SCC), which was established in 1972 specifically for that purpose. Proceedings are confidential (unless both parties agree otherwise) and the commission meets at least twice a year. It is credited with having resolved a number of SALT I compliance issues under the Nixon, Ford and Carter administrations.

The SCC's stock seems to have fallen recently, and the commission would clearly benefit from certain improvements. On the other hand, compliance diplomacy need not be exclusively conducted by means of a standing mechanism. Parallel diplomatic channels, which are free of procedural constraints and are responsive to the immediate concerns of top decision makers, might also be needed. Just such a "back channel" between then Secretary of State Kissinger and Soviet Ambassador Anatoly F. Dobrynin supplemented the formal SALT I negotiations. A well-conceived, alternative line of direct communication, designed to iron out the most pressing compliance problems, could be a useful complement to the SCC.

Above all, arms-control agreements should be placed in proper perspective. They are, like the arms race itself, primarily self-serving tools of national interests. An airtight formulation and an unimpeachable implementation are not realistic goals for any arms-control agreement and should not serve as criteria by which such agreements are evaluated. The benefits that arms-control treaties hold for international security and governmental budgets can be assessed only in the light of the alternatives their absence would create.

# The Dynamics of Proteins

*The molecules essential to life are never at rest; they would be unable to function if they were rigid. The internal motions that underlie their workings are best explored in computer simulations*

by Martin Karplus and J. Andrew McCammon

The study of how proteins serve the needs of a living organism is a curious case in which a method that yielded dramatic advances also led to a misconception. The method is X-ray crystallography, by which the structure of protein molecules in a crystal of the substance is determined from the way the crystal diffracts, or scatters, a beam of X-radiation. The intrinsic beauty and the remarkable detail of the structures obtained from X-ray crystallography resulted in the view that proteins are rigid. This created the misconception, namely that the atoms in a protein are fixed in position. Most attempts to explain protein function (for example, how proteins act as enzymes, or biological catalysts) have been based on the static structures. The specificity with which an enzyme binds a particular molecule, called its substrate, and the efficacy with which an enzyme catalyzes a particular biochemical reaction have been presumed to originate from complementary interactions between a substrate and a rigid enzyme—interactions that were likened to the fitting of a key into a lock. Cases in which the conformation, or shape, of a protein was known to change, as when the protein hemoglobin binds oxygen, were generally treated as abrupt transitions between otherwise static structures.

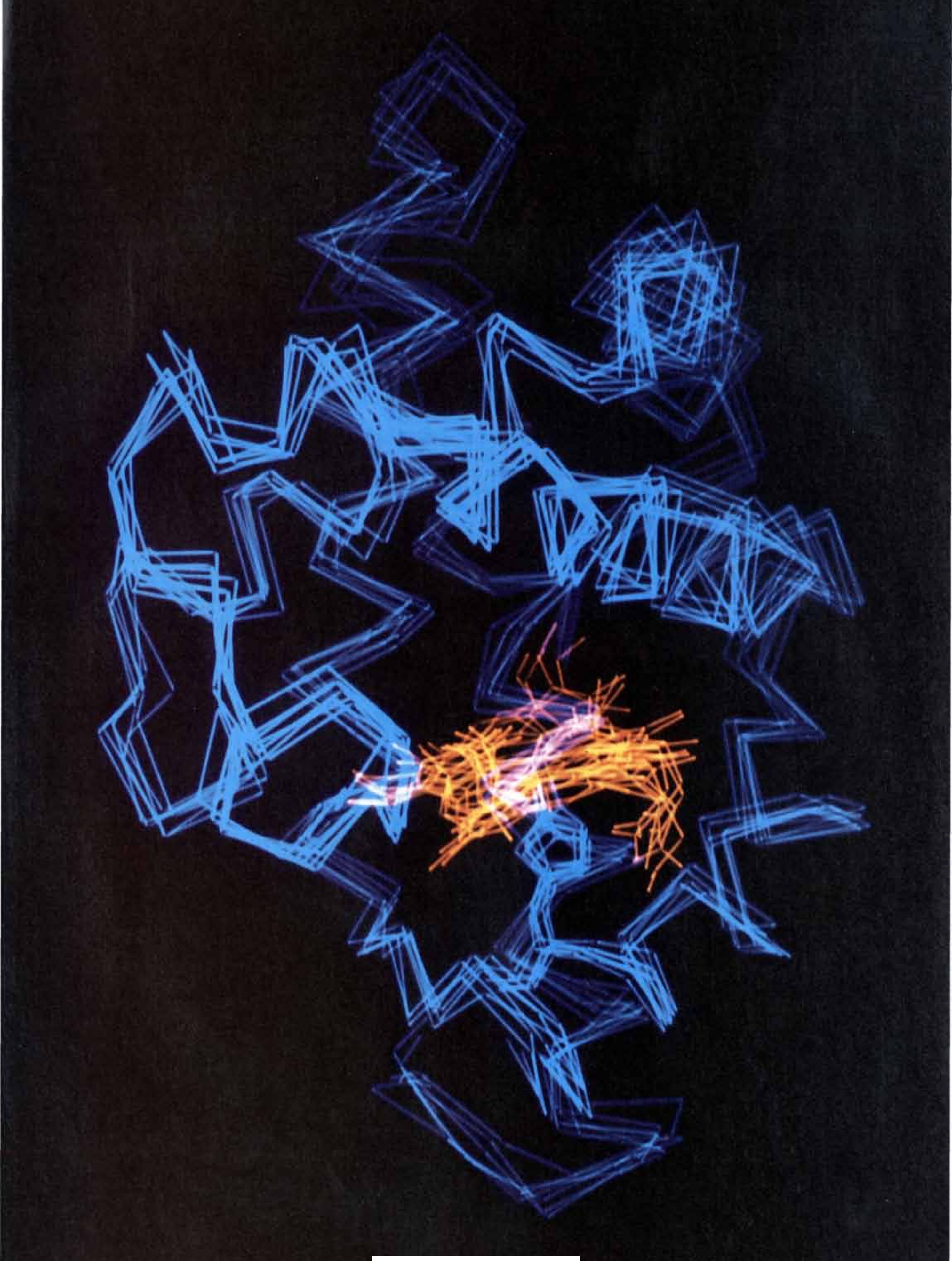
In the past 10 years this static view of proteins has been undergoing a fundamental revision. It is now recognized that the atoms in a protein molecule are in a state of constant motion, so that what the crystallographer finds in a crystal is at best a representation of a protein's average structure. The atoms in each protein molecule exhibit sizable high-frequency fluctuations about this average. In brief, a protein is dynamic: it is constantly changing the details of its conformation. A crystal prepared for X-ray analysis may consist of some $10^{20}$ protein molecules, yet it is highly improbable that at any instant of time even one individual protein molecule has the average structure.

To be sure, the knowledge of the average atomic positions in a protein remains immensely valuable: it serves as a point of departure for attempts to produce more nearly complete descriptions of protein activity. The dynamic picture makes it possible to understand phenomena that cannot be explained by the static model. The descriptions of the dynamics are derived from many sources; experimental techniques, including X-ray crystallography, make important contributions. Theoretical approaches, however, have uncovered details of protein motions far beyond what one could ever hope to find by means of experimental measurements. In some cases computer simulations of the atomic motions in a protein have shown that the activity of the protein would be impossible if the molecule were fixed in its average structure.

The specific structure and fluctuations of proteins enable them to execute many of the tasks required for an organism to function. For one thing, all enzymes are proteins; that is, proteins are the catalysts that speed up the essential reactions of living systems (including the synthesis of proteins themselves) so that the reactions proceed at the required rates. Other proteins serve to transport small molecules, electrons and energy to the parts of the organism where they are needed. Many proteins have structural roles; for example, they are constituents of fibrous tissue and muscle.

Because of their varied functions, many different proteins are required. A single-cell organism, the well-studied bacterium *Escherichia coli,* is known to have 3,000 different proteins, with many copies of some, so that one bacterium may include about a million individual protein molecules. Each protein is made up of a specific number of small units, the various amino acids. Each of the 20 different amino acids is characterized by a side chain, a distinctive chemical group that ranges in complexity from a hydrogen atom in the simplest amino acid, glycine, to elaborate rings of atoms in the most complex amino acid, tryptophan. In a protein the amino acids are linked together in a linear array called a polypeptide chain. Proteins commonly consist of from 50 to 500 amino acids, which corresponds to some 500 to 5,000 atoms. The precise sequence of amino acids determines the average structure and other properties of the protein. In particular, the balance of the attractive and repulsive forces between the individual atoms of which the amino acids are composed

**MOTIONS OF A PROTEIN on a picosecond (trillionth of a second) time scale are illustrated by superposing seven positions of a myoglobin molecule seen at intervals of five picoseconds in a molecular-dynamics simulation implemented on a computer. Like all other proteins, myoglobin consists of amino acids linked in a polypeptide chain (*blue*). The chain is shown in a simplified version that includes only the position of the central carbon atom of each amino acid. The chain is folded into a characteristic three-dimensional conformation, or shape; in myoglobin the conformation consists of eight segments called alpha helixes, connected by short loops each made up of a few amino acids. The helixes move in the simulation but retain their shape. The myoglobin molecule also contains a complex organic molecule called the heme group (*orange*). Myoglobin stores oxygen in muscle tissue by binding an oxygen molecule to the iron atom at the center of the heme group. The computer-graphics display, and others accompanying this article, were generated by John Kuriyan of Harvard University and the Massachusetts Institute of Technology by means of the computer program HYDRA, written by Roderick E. Hubbard, now at the University of York.**



causes the chain of a globular protein to fold in the characteristic way essential to its fluctuations and its functions.

The first step toward understanding the role of atomic fluctuations in protein function is to determine the nature of the fluctuations themselves. This includes the study of their magnitude (how large they are), their probability (how often they occur) and their time scale (how long they take). The most direct approach to protein dynamics is to treat each atom in the protein as a particle responding to forces in the way prescribed by Newtonian physics, in accord with Newton's equations of motion.

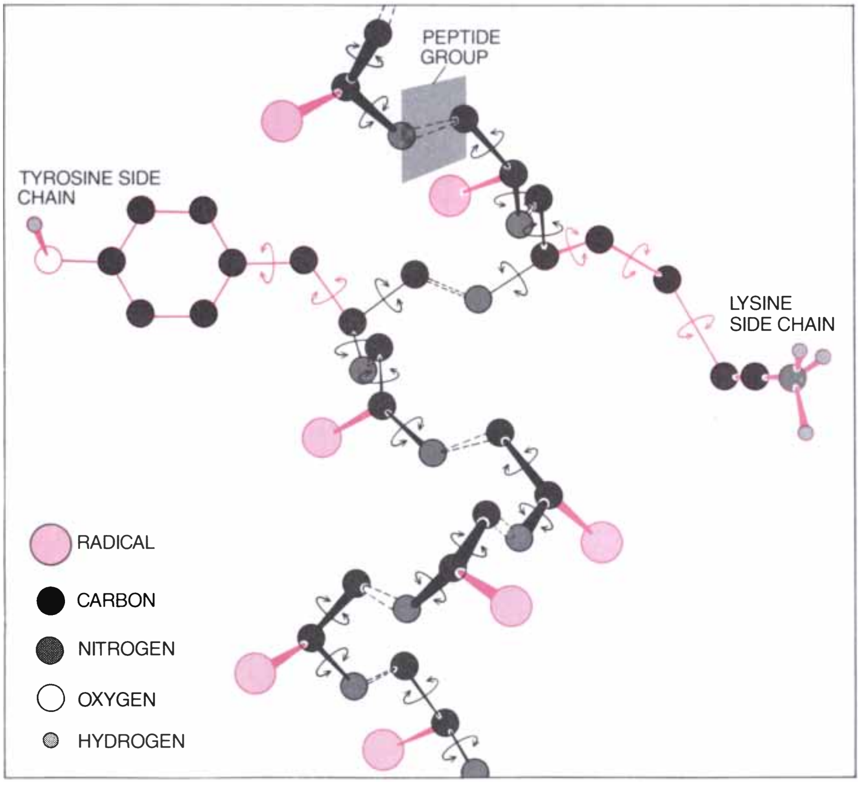
A simple analogy is the calculation of the motion of a pendulum or a particle suspended by a spring. Given the position and velocity of the particle and also the magnitude and direction of the force acting on it at a particular instant, the position and velocity at a slightly later time can be obtained by extrapolation, or integration of the equations of motion. The results are accurate if the extrapolation interval, called the time step, is small enough so that the force at the new position is not very different from the force at the original position. By repeating the extrapolation procedure many times (that is, by integrating Newton's equations of motion over a longer time period), the trajectory of the particle is obtained, in the form of a list of positions and velocities at successive times, one position and velocity for each time step.

The chemical bonds between atoms along the polypeptide chain in a protein are very much like springs. There are also weaker forces between nonbonded atoms, however, including forces that prevent more than one atom from occupying the same point in space at any one time. Thus in a protein consisting of many atoms the total force acting on any one atom at any given time depends on the positions of all the others. Not surprisingly, the solution of Newton's equations of motion for determining the positions and velocities of all the atoms in a protein requires a high-speed computer. Such calculations constitute what is called a molecular-dynamics simulation.

To begin the simulation one needs a representative set of initial positions. These are derived from the atomic positions available from X-ray-crystallographic data. The crystallographic positions cannot themselves serve as an initial structure. Since they correspond to an average structure, some of the atoms are placed in positions not typical of any actual structure (for example, two atoms may be too close together). This structural distortion can in turn result in large, unrealistic forces on some of the atoms.

A procedure called dynamic equilibration overcomes this problem and prepares the protein in a state with positions and velocities appropriate for the start of its calculated trajectory. In the equilibration procedure the crystallographic positions of the atoms of the protein (and of surrounding solvent molecules one may want to include in the simulation) are adjusted to relax the unrealistic forces. Then small random velocities corresponding to a given temperature (near absolute zero) are assigned to the atoms. (The temperature of a molecule is a measure of the magnitude of its average atomic velocities.) With this set of positions and velocities and the values of the instantaneous forces, the atoms are allowed to move in accord with Newton's equations of motion for 100 time steps or so. Sets of successively greater random velocities are then assigned and again the atoms are allowed to move, until the desired temperature (room temperature, for instance) is attained; the equilibration is completed by having the system evolve spontaneously for a period of time by integrating the equations of motion until the average temperature and structure remain stable.

The molecular-dynamics simulation takes the equilibrated protein as its starting point. Since the forces between atoms in the protein must change only slightly in each successive step of the simulation, the time steps must be short compared with the time scale of the fastest motions in the protein. A typical time step is on the order of $10^{-15}$ second (one femtosecond). The calculated positions and velocities of the atoms are stored on magnetic tape for subsequent analysis and graphic display. In a typical simulation the trajectory traces the dynamics of the protein through a period of about $10^{-10}$ second (100 picoseconds); newer, faster computers make simulations spanning $10^{-9}$ second (one nanosecond) or even longer feasible. (The latter, which require as many as a million time steps, can consume several hundred hours of computational time on a supercomputer.) Fortunately many interesting motions of a protein are fully developed in 100 picoseconds or less.




**SITES OF FLEXIBILITY in a polypeptide chain enable the chain to fold into the conformation characteristic of the protein; the sites also facilitate the fluctuations the protein atoms make with respect to their average positions. The drawing depicts only the principal atoms of a polypeptide chain. The backbone of the chain (*black bonds*) consists of carbon and nitrogen atoms; the linkages called peptide bonds are rigid, whereas the intervening bonds allow rotations (*curved arrows*). A side chain (*colored bonds*) is linked to each amino acid's central carbon atom. The side chains shown in detail also contain rotatable bonds.**

Even in that short time the amount of detail stored in a trajectory is prodigious; in effect the computer puts the researcher inside the protein.

The motions of the atoms in a protein, as revealed by molecular-dynamics simulations, tend to share certain characteristics that can be explained in terms of the basic structure of proteins. Like most polymers, the polypeptide chain that makes up a protein is flexible because many of the strong bonds holding pairs of atoms together are free to twist. The twisting, or bond rotation, allows one part of the polypeptide chain to move with respect to another. As the polypeptide chain twists and turns, its various side chains move with it. The side chains themselves have bonds around which rotations can occur, imparting additional flexibility.

The flexibility of the polypeptide backbone and of the side chains is what enables each globular protein to fold into its characteristic native, or average, structure. Even in the folded protein, however, the thermal energy corresponding to the atomic velocities at room temperature is sufficient to allow twisting motions. Such motions are the primary source of the atomic fluctuations in the protein. Since a globular protein has its atoms packed close together, the atomic fluctuations are restricted. Nevertheless, the combination of many small local motions can produce large-scale, more global displacements of one part of a protein with respect to another.

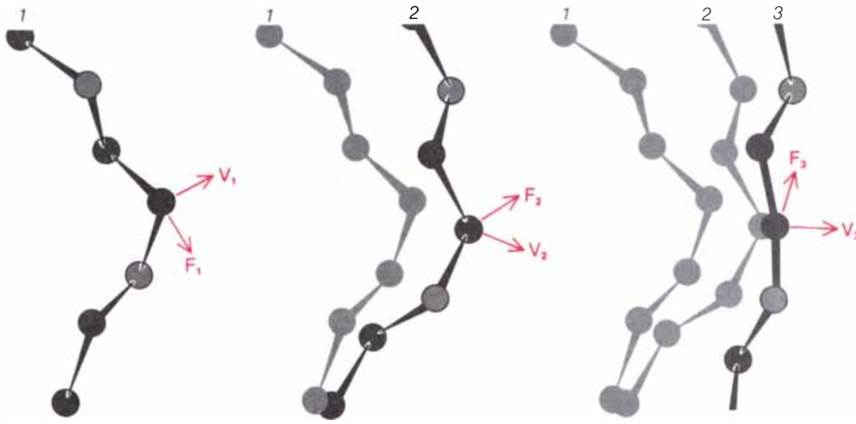
Over short periods, say intervals of a few tenths of a picosecond, the most prominent fluctuations within a protein are the local motions in which peptide and side-chain groups typically rotate through angles of from 20 to 60 degrees around the covalent bonds linking the groups to the rest of the polypeptide chain. There is usually little coherence in these motions. Rotating groups are likely to collide with neighboring groups, so that the motion of each group is frequently interrupted. The motion of each group within the protein is therefore somewhat like that of a molecule in a liquid, where collisions with solvent molecules slow down the molecule (frictional drag) and produce an erratic trajectory associated with a "random walk," or diffusive Brownian motion.

Over longer periods of time, say several picoseconds, the important features of motion become less like those in a liquid and more like those of a solid. In particular, the forces maintaining the native structure of the protein do not allow individual atoms to move very far from their average positions. The solidlike motions tend to be more collective than the local oscillations. They range from simultaneous displacements of a few neighboring amino acids to deformations of the entire molecule. These motions are damped to varying degrees by frictional effects arising from collisions between protein atoms and between protein atoms and solvent molecules. Nevertheless, the collective motions that develop over periods of several picoseconds tend to make the dominant contribution to the overall amplitudes of the atomic fluctuations in a protein. The reason is straightforward: owing to the tight packing of the protein's interior, a typical atom cannot move far unless its neighbors move also. The largest motions occur at the surface of the protein, where the packing constraints are less important. Individual atoms in the interior of the protein may have fluctuations of up to half an angstrom unit, a distance comparable to the radius of an atom. Atoms at the surface may make excursions as great as two angstroms.

One globular protein in which molecular dynamics is important to the molecule's function is myoglobin, the protein that stores oxygen in muscle tissue. In whales, for example, a large supply of myoglobin provides the oxygen required throughout the long periods when the animal is submerged. Myoglobin is the first protein whose structure was determined, by John C. Kendrew of the University of Cambridge. The myoglobin he subjected to X-ray-crystallographic analysis came from sperm whales whose meat served as food in England during World War II.

In general terms the workings of myoglobin are well understood. The oxygen molecule stored in myoglobin binds reversibly to the iron atom at the center of a flat, complex organic molecule called the heme group. The heme group lies deep in the myoglobin molecule, so that the surrounding, densely packed globular protein protects it from water. (The water would oxidize the heme group's iron atom from its ferrous state, $Fe^{+2}$, to its ferric state, $Fe^{+3}$, in which it cannot bind oxygen.) The packing, however, is so dense that the oxygen cannot penetrate to the heme. If the atoms in myoglobin were fixed in the positions found in the X-ray-crystallographic structure, myoglobin would be useless: the time required for an oxygen molecule to bind to the heme group or to get out again when needed would be much longer than a whale's lifetime.

From what we have already said about atomic fluctuations in proteins it is possible to guess how the problem is solved: the energy barriers preventing oxygen from getting in and out of the rigid protein are lowered by a combination of local motions. To explore this possibility conceivable paths to the heme group's pocket in myoglobin have been examined and the barrier along these paths calculated both for the rigid protein (that is, a myoglobin molecule with the structure obtained from X-ray crystallography) and for a more realistic model in which the structure can relax as the oxygen pass-




**MOLECULAR-DYNAMICS SIMULATION gives a detailed description of the motions in a protein by treating the atoms composing the protein as Newtonian particles. The bonds between the atoms are treated much like stiff springs; atoms that are not bonded together interact through weaker forces. A simulation consists of a series of many small steps ($10^{-15}$ second in length) in which the position and velocity of each atom are calculated. Given the positions and velocities at a time $t_1$ and a method for determining the magnitudes and directions of the forces acting on the atoms at that time, the positions and velocities after a time step ($t_2$) are obtained by solving Newton's equations of motion. Calculations of the forces at $t_2$ make it possible to determine positions and velocities at time $t_3$. Successive recalculations yield the internal motions of a protein over periods as long as $10^{-9}$ second.**

es through the protein matrix. For all the paths investigated the rigid protein resulted in energy barriers of at least 100 kilocalories per mole. For barriers of that magnitude the time required for an oxygen molecule to enter or exit would be many billions of years.

To simulate the possible fluctuations that might open paths for oxygen to get into or out of myoglobin, the motions of side chains that produce the high energy barrier were considered. For one particular path the side chains of three amino acids (designated histidine *E*7, threonine *E*10 and valine *E*11) seemed to be the dominant obstacles. By simulating the rotation of each of these side chains the energy required to move them out of the way has been determined. An energy of 8.5 kilocalories per mole suffices to open the path so that the energy barrier for the oxygen is decreased to about 5 kilocalories per mole. The total energy required is therefore reduced from 100 kilocalories per mole to about 14, an amount on the order of magnitude suggested by experiments that actually measure the energy barrier to oxygen binding.

Finally, in order to examine the oxygen motion in detail, a series of molecular-dynamics trajectories was calculated for oxygen molecules starting at the heme. Because each simulated trajectory could be followed for only a limited time, not all the oxygen molecules escaped; some remained in a pocket near the heme and others ended up somewhere else inside the protein. Still, many did reach the outside. In a typical path the oxygen molecule encountered a series of obstacles, each with its own energy barrier. The oxygen would spend a long time in a given energy "well," moving about and colliding with its walls. Then, when a protein fluctuation significantly lowered the barrier of the well or the oxygen was energized by collisions with the protein atoms, or when both happened simultaneously (the likeliest possibility), the oxygen would move rapidly over the barrier and into the next well, where the process would be repeated.

The complexity of each path and the number of possible paths make it likely that the motion of the oxygen through the protein is diffusive in character. Once the oxygen crosses a barrier it does not necessarily move on to the next one. Its direction may be randomized by collisions and it may recross the barrier in the wrong direction. The resulting behavior is a random walk similar to that of a solute molecule diffusing in solution.

The analysis of myoglobin suggests that what happens in this globular protein is of general significance. Ligands, the molecules that bind to many proteins, may be unable to enter or leave the protein if the atoms in the protein are constrained to occupy their average positions. In enzymes, therefore, the fluctuations of side chains or other atomic constellations may often be a prerequisite for substrates to enter and bind and for reaction products to leave.

Many important steps in the function of a protein take place over relatively long periods of time: they require periods of nanoseconds ($10^{-9}$ second) to milliseconds ($10^{-3}$ second) or even longer. For example, the rate of the reaction catalyzed by some enzymes is limited by the time needed for certain groups of atoms in the enzyme to shift from one conformation to another, in which they can participate in the catalysis. The actual shift, when it occurs, is fast; the long time scale arises because the atomic groups involved in the transition become activated only rarely. That is, the fluctuations that make the transition possible are infrequent occurrences.

A conventional molecular-dynamics simulation is of little use in the study of activated processes; the infrequent occurrence of the conformational transition makes it highly improbable that it will take place in the interval spanned by the simulation. Yet if one knows roughly what displacements are involved in the transition, it becomes possible to study the process in detail by means of a specialized simulation method.

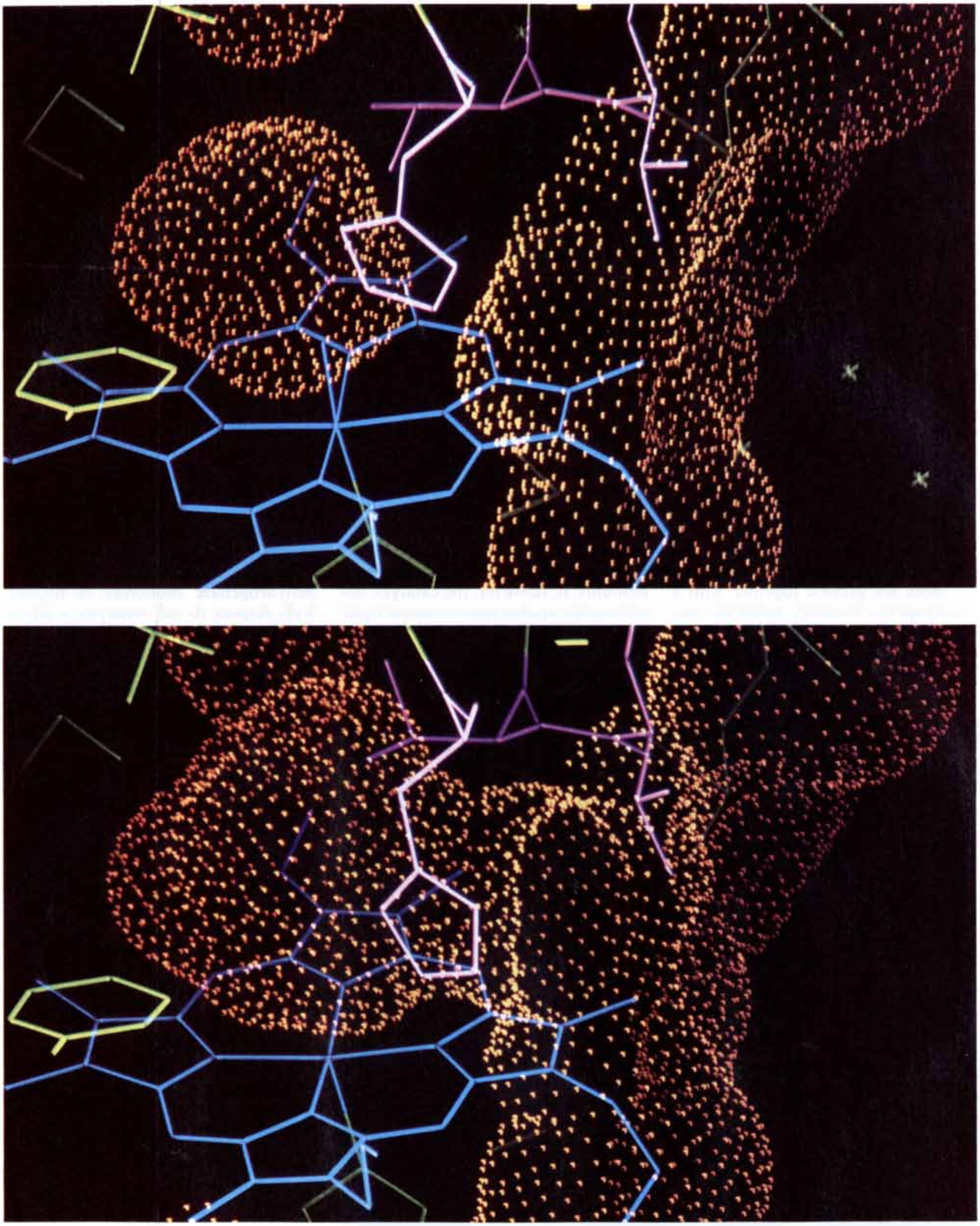
Such simulations are carried out in two steps. First the magnitude of the energy barrier to the transition is calculated from a sequence of simulations; these are of the conventional type except that the simulated molecule is constrained to move only within a succession of limited regions that take it along the path of the transition. Then trajectories are calculated that begin with the molecule in configurations at or near the top of the barrier. In this way the need to wait for the chance occurrence of an activated molecule is eliminated. The trajectories that emerge from these simulations, together with the calculated magnitude of the energy barrier, enable one to predict the rate of occurrence of the transition.

The application of this activated-dynamics method to some simple processes has yielded insights into the nature of the local fluctuations involved in activating a conformational transition. The transition we shall describe (more as a representative of an activated process than as a transition that has an important biological role) is the rotation of a tyrosine ring deep inside the globular protein called bovine pancreatic trypsin inhibitor. As its name implies, trypsin inhibitor can arrest the activity of the digestive enzyme trypsin by binding to the active site at which trypsin catalyzes the cleaving of protein molecules. For its part, the side chain of the amino acid tyrosine includes six carbon atoms that form a flat, hexagonal ring. The rotations of the ring are of interest because they have been studied experimentally, by nuclear magnetic resonance. The technique of molecular-dynamics simulation shows that the rotation has a collective character in that it is preceded by a displacement of a section of the polypeptide chain that lies above one face of the ring.

Displacement of the chain makes a critical contribution to the rotation mechanism. It moves atoms out of the path of the ring, resulting in a substantial reduction in the energy barrier impeding the rotation. The energy required for the backbone distortion is much smaller than that arising from the repulsive contact between the ring atoms and the undistorted backbone. Furthermore, the collective fluctuation helps to initiate the rotation. It opens a small volume into which the ring can rotate in response to the collisions it makes with the remaining neighbor atoms. The resulting bias in favor of rotation allows the ring to accumulate the rotational kinetic energy necessary for surmounting the remaining energy barrier.

In this way the occurrence of a simple conformational transition (the tyrosine rotation) is regulated by collective motions within the protein. Given the close packing of the atoms in globular proteins, such transitions are likely to be a common occurrence. Since their rates are sensitive to relatively large-scale features of protein structure, the transitions may provide a way of regulating a protein's activity. For example, the binding of a small molecule at one site on a protein may alter the motions of the protein so as to change the kinetics of a process at a second site.

An example of a protein where large-scale motions are involved in enzyme activity is provided by liver alcohol dehydrogenase, the enzyme that oxidizes ethyl alcohol and thus helps to limit the intoxication brought on by drinking wine or other spirits. (People who lack the enzyme are abnormally sensitive to alcohol.) The true role of the enzyme is thought to be the catalysis of some other biochemical reaction, since evolution would seem to

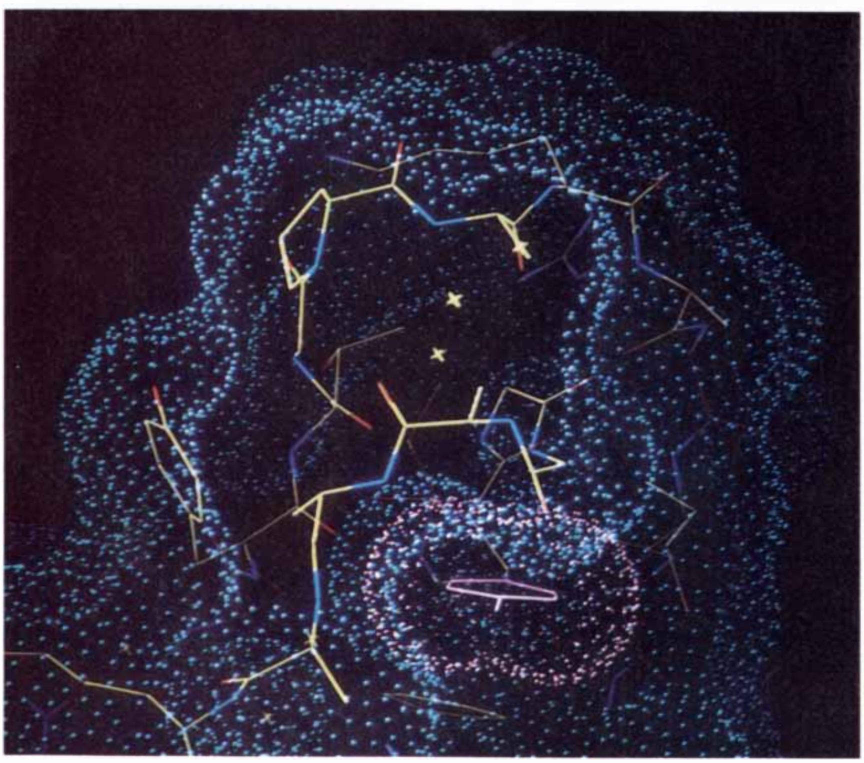


**CHANNEL OF ESCAPE** for an oxygen molecule bound to the heme group of a myoglobin molecule is opened by the rotations of three amino acid side chains. The top image shows the static structure of myoglobin determined by X-ray crystallography. There is no path from the surface of the molecule (*orange dots*) to a pocket inside the molecule just above the heme group (*blue*). The pocket is near the site where the heme group's iron atom binds an oxygen molecule. In the bottom image the side chains of three amino acids (*purple*) have been rotated. As a result an escape channel opens. The positions of the other backbone and side-chain atoms (*green*) are unchanged in this simulation. If the structure of the myoglobin molecule were rigid, so that the rotations of side chains were impossible, an oxygen molecule might take billions of years to enter or leave the binding site. Detailed simulations suggest that an oxygen molecule encounters a succession of barriers, so that it moves in a "random walk" much like that of a small molecule in a liquid.

have had no reason to produce a dealcoholizing enzyme. In any case, liver alcohol dehydrogenase exemplifies the enzymes and other proteins, such as immunoglobulins, that consist of two or more domains connected by strands of polypeptide chain, which may serve as hinges. In many enzymes the catalytic region lies between two globular domains, so that access to the site is regulated by hinge-bending deformations of the entire protein molecule—deformations that separate the lobes or bring them together. The rate at which substrates are bound or products are released may therefore depend on the dynamics of the domains.

Liver alcohol dehydrogenase is a dimer, a molecule consisting of two identical monomers, or subunits. In turn each monomer has two lobelike domains. X-ray crystallography by Carl Brändén, Hans Ecklund and their colleagues at the University of Uppsala reveals that the monomer has two configurations. There is an open structure, the so-called apoenzyme, in which the lobes are spread apart, and a closed structure, the holoenzyme, in which the lobes are pressed together with a coenzyme (an ancillary molecule that participates in an enzymatic reaction) bound to one of the lobes. In this case the coenzyme is the electron-transporting molecule called NADH. The open structure provides a way for the coenzyme to get to its binding site, whereas the closed structure provides an environment in which the substrate and the coenzyme are protected and the reaction can proceed efficiently (the second lobe is the enzyme's catalytic domain). When the reaction has taken place, an opening fluctuation moves the domains apart, the reaction product escapes and the enzyme is ready to go through another reaction cycle.

As a starting point for the examination of the opening and closing fluctuations, it was assumed that the two domains of a monomer of liver alcohol dehydrogenase move like rigid bodies connected by a flexible hinge. The assumption can be tested by superposing the open and the closed crystal structures (that is, the apo and the holo configurations). When the coenzyme-binding domains of the apoenzyme and the holoenzyme are made to coincide, the positions of the atoms in the catalytic domains differ by large amounts. If, however, the catalytic domain in the apoenzyme is rotated rigidly so that the hinge between the lobes closes by about 10 degrees, most of the atoms in the apoenzyme catalytic domain become superposable on the equivalent atoms of the corresponding domain in the holoenzyme.

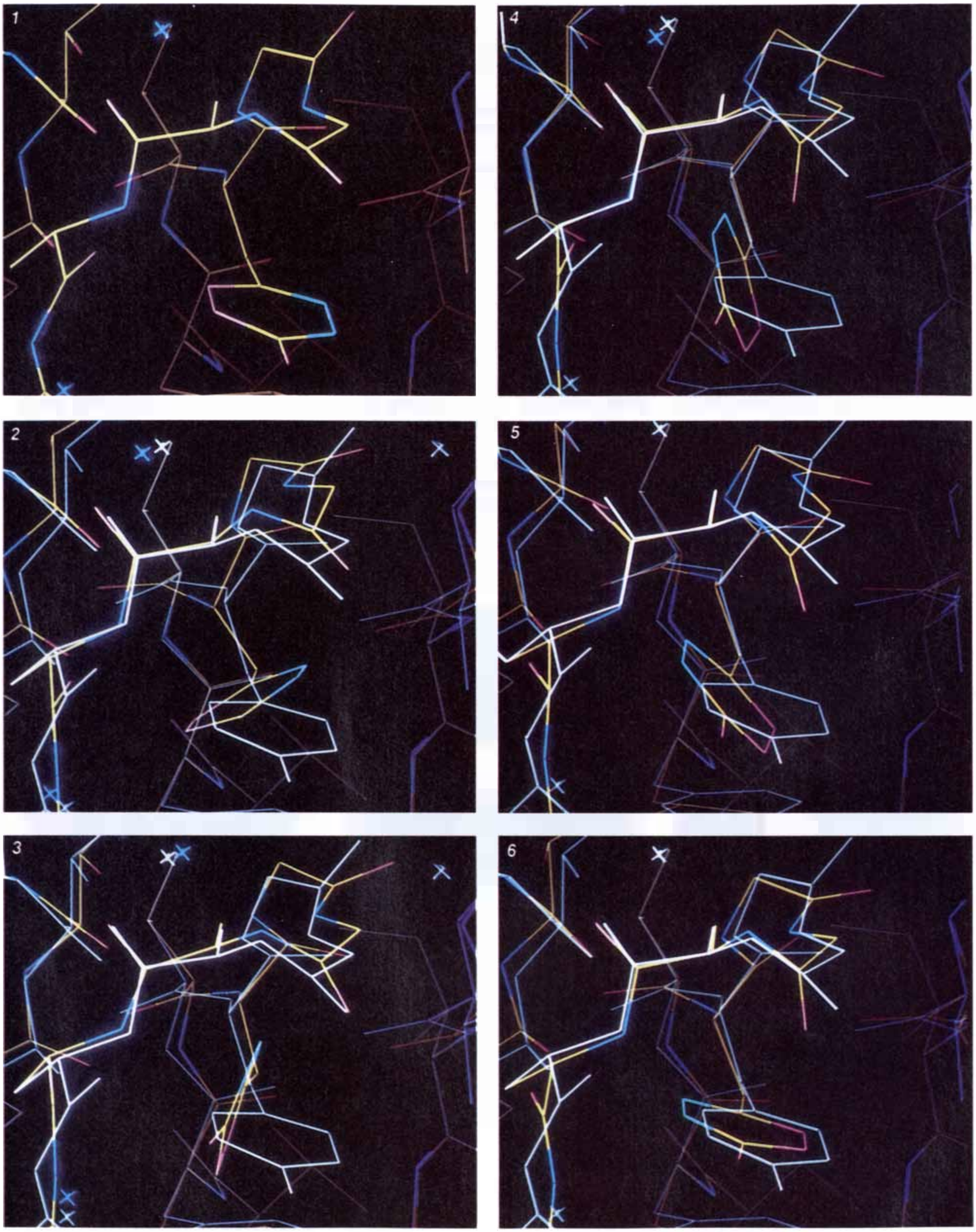


**HEXAGONAL RING of carbon atoms characterizes the side chain of the amino acid tyrosine (*magenta*), seen deep in the protein called bovine pancreatic trypsin inhibitor. The image shows the bonds that establish the ring (*magenta lines*) and in addition the van der Waals surface of the ring (*magenta dots*), which fits tightly into the protein's static structure (*blue dots*). In the rigid protein the tyrosine ring could not flip, or rotate by 180 degrees.**

Nevertheless, the assumption of rigid rotation is a gross oversimplification. A model incorporating the types of interaction required to calculate the forces for a molecular-dynamics simulation shows that the holo-like configuration produced by rigid rotation of the apo catalytic domain has an energy several thousand kilocalories per mole greater than that of the apo structure. This energy corresponds to an immense barrier with respect to the available thermal energy—a barrier that would prevent any rotation. The barrier is introduced by neglecting the small, local atomic fluctuations shown to occur by molecular-dynamics simulations. For hinge bending to be possible at normal temperatures it is essential that unfavorable atomic interactions generated by the rotation be allowed to relax.

One way of simulating a relaxed hinge-bending motion is by an adiabatic potential calculation. In this procedure one domain in the liver alcohol dehydrogenase monomer is rotated with respect to the other by a small amount, say one degree. The atoms in the hinge region and in areas of contact between the two domains are then allowed to undergo positional fluctuations that minimize the energy of the rotated structure. A rotation by another one-degree increment is then introduced, and the minimization is repeated. The final result, accumulated from a sequence of incremental rotations and relaxations, is an evaluation of the potential energy of the molecule as a function of the hinge-bending rotation angle. The energy curve covering the range from the apoenzyme to the holoenzyme turns out to be rather flat; in fact, the random Brownian motions of the lobes of the molecule in solution at room temperature would have enough energy to produce spontaneous closings and openings, on a time scale of nanoseconds.

The atomic displacements that constitute the relaxation of the monomer are, with few exceptions, smaller than .5 angstrom; they are thus comparable in size to the normal room-temperature atomic motions in a protein. They take place, as described above, on a picosecond time scale and are therefore far more rapid than the large-scale hinge-bending motion itself. This justifies the simulation technique: each change in the rotation angle allows ample time for the local relaxations that are introduced by energy minimization. The analysis of liver alcohol dehydrogenase (and of several other hinge-bending proteins that have been




**FLIP OF THE TYROSINE RING** in the interior of pancreatic trypsin inhibitor exemplifies molecular processes in which random fluctuations in a protein molecule make possible a sudden conformational transition. The field of view in this molecular-dynamics simulation includes the ring and also a part of the polypeptide backbone that makes contact with the upper face of the ring, as is shown in the preceding illustration. The ring is colored to aid in viewing its rotation: its left edge is red and its right edge green. Its initial position (*1*) is repeated for reference in subsequent steps of the simulation (*white hexagon in 2–6*). Early in the simulation the backbone lifts away from the tyrosine ring. This backbone fluctuation opens space in which the ring can rotate; the ring's collisions with its remaining neighbor atoms tend to drive the rotation. The images span a time of two picoseconds, during which a full 180-degree flip takes place.

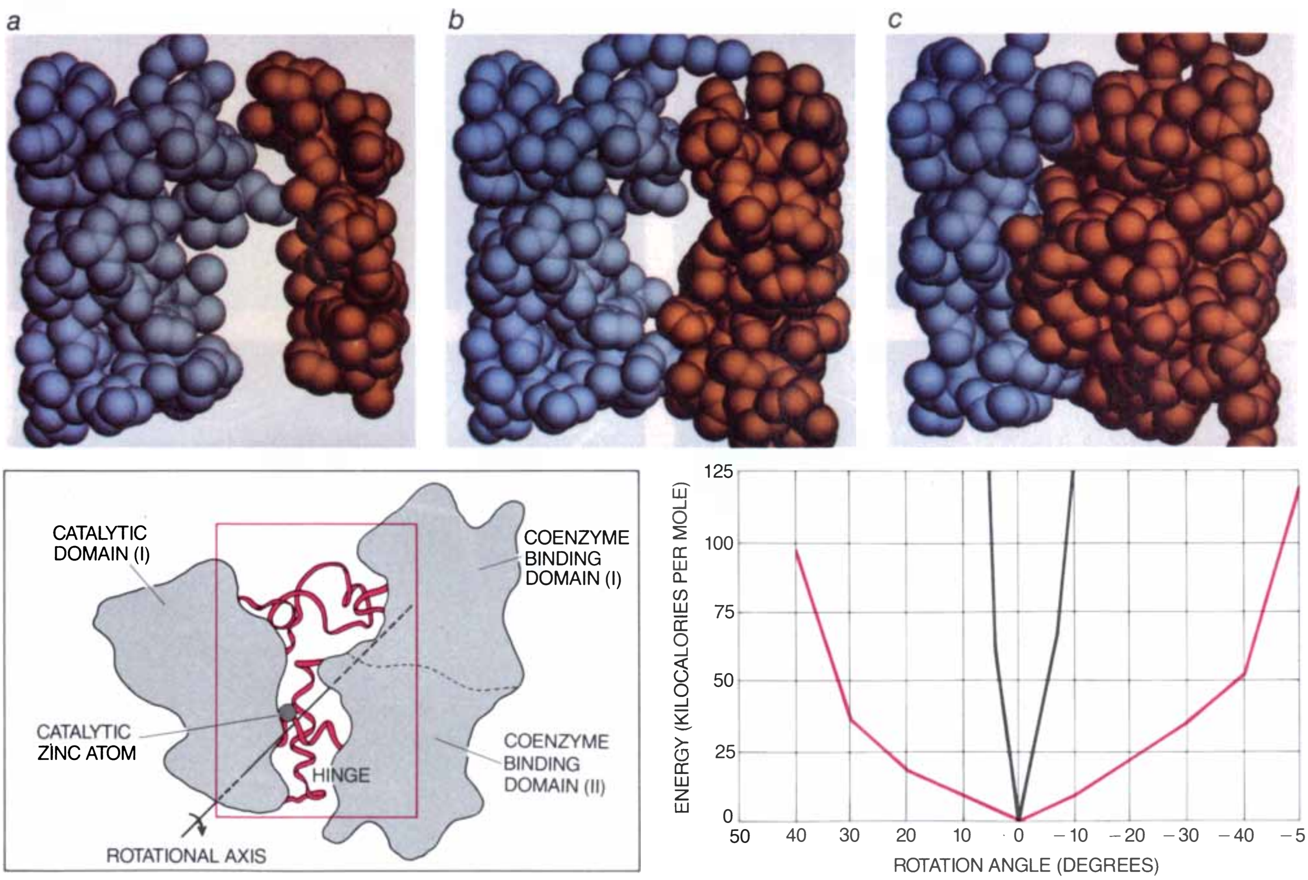
studied theoretically) makes clear the important role of small, high-frequency fluctuations in facilitating some larger and more collective motions of proteins.

One of the objectives of molecular-dynamics simulations is to provide a detailed and quantitative description of how enzymes enhance the rates of biochemical reactions. Enzymes are thought to act by a combination of mechanisms. By binding the reactants they bring them together; moreover, enzymes can place charges associated with the side chain of certain amino acids in positions to aid the reaction, and so on. Still, no enzymatic reaction is understood well enough so that its rate can be predicted. The motions of the atoms composing the enzyme, the reactants and even the solvent water molecules at and near the active site of the enzyme are all likely to have a significant influence on the rate. In many enzymes the active site remains accessible to solvent throughout the reaction and water itself acts as one of the reagents; in other enzymes (such as liver alcohol dehydrogenase) the solvent tends to be excluded.

In enzyme catalysis, as in chemical reactions generally, atomic motions clearly have an essential role. An approach based on molecular dynamics should hence be very fruitful. The computer simulation of an entire enzyme and its solvent would be an inefficient (and expensive) way to study the localized dynamics directly involved in a catalyzed reaction. Instead the molecular-dynamics simulation can be restricted to a "reaction zone" that includes the active site and its neighborhood, the substrate and the nearby solvent molecules. The rest of the system cannot be ignored. It is replaced by a boundary region in which solvent molecules are kept from escaping by a repulsive force and protein atoms are constrained by forces that result in their having the fluctuations they would have if the simulation included the entire protein. Moreover, the equations of motion for the protein atoms and solvent molecules in the boundary region are modified to include Brownian-type coupling terms that allow these atoms and molecules to absorb or release energy as if they were surrounded by the rest of the system.

The boundary simulation method is now being applied to a number of proteins, including the enzyme ribonuclease, which cleaves ribonucleic acid (RNA) by cutting between two nucleotides, the units that make up a strand of RNA. Indeed, simulations have been done for the native enzyme (that is, for ribonuclease with no substrate nucleotides bound to the enzyme) and for various intermediates along the reaction path by which the enzyme cleaves an RNA strand. One finding that has emerged from the calculations concerns the stability of the parts of the structure of ribonuclease that are involved in binding the substrate and by that token are important in determining the specificity of the enzyme.

The simulations have shown, for example, that a number of amino acids in ribonuclease make a network of hydrogen bonds to the substrate; hydro-

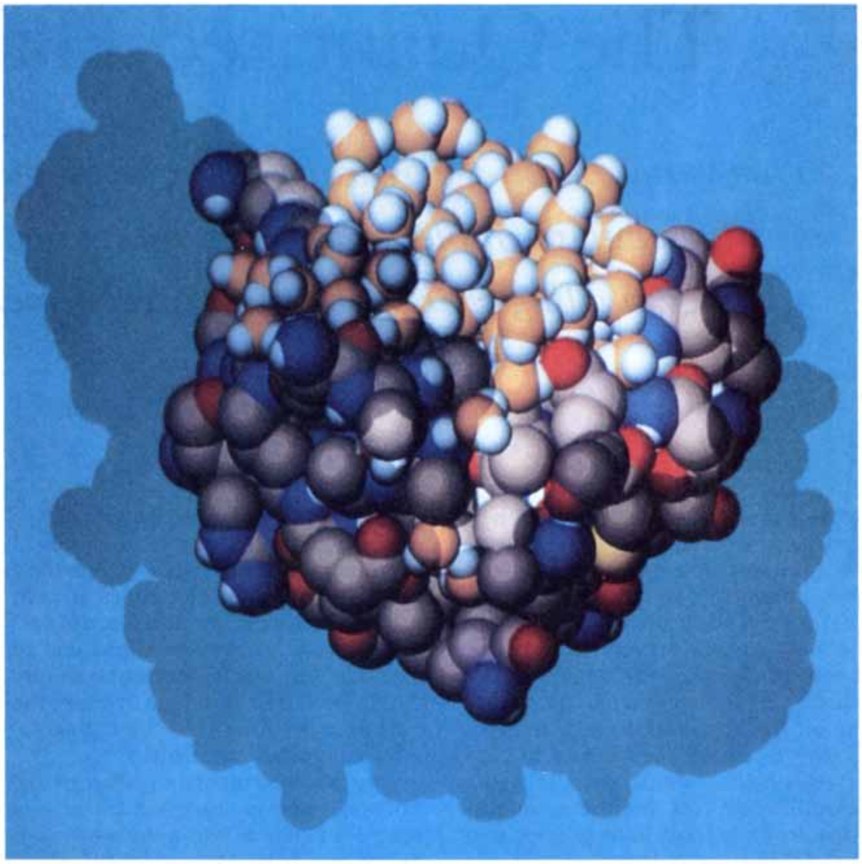


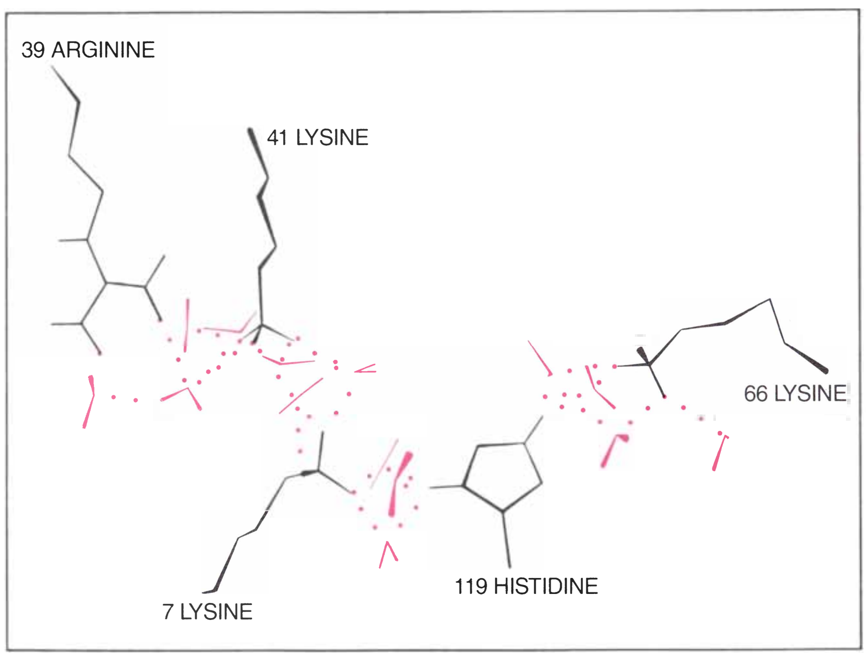

**HINGE-BENDING MOTION of the enzyme called liver alcohol dehydrogenase brings two lobes of the molecule together, providing a protected environment for the chemical reaction catalyzed by the enzyme. Simulations of the parts of the lobes that meet are shown, in computer graphics prepared by Carl Bränden of the University of Uppsala, for a hinge angle of 30 degrees (*a*), in which the hinge is open, 0 degrees (*b*) and −40 degrees (*c*), in which the hinge is closed. The angle of 0 degrees places the enzyme in the apo, or relatively open, conformation observed by X-ray crystallography. The simulations establish that if the hinge bending were rigid (*black curve in bottom chart*), the bending motions would require much more energy than the amount available at room temperature. If, however, the atoms in the molecule fluctuate during the bending (*colored curve*), the hinge-bending motions become quite possible.**

gen bonds are electrostatic linkages between pairs of charged or partially charged (polar) atoms that share a hydrogen atom. Some of these networks involve the side chains of asparagine 67, glutamine 69 and asparagine 71 (which form the enzyme's binding site for the nucleotide called adenine) and threonine 45 and serine 123 (which form the binding site for the nucleotide called cytosine). In the absence of a substrate, water molecules occupy the positions of the polar substrate atoms in the binding sites and produce a network of hydrogen bonds that mimic the substrate interactions and stabilize the positions of the binding-site amino acids. Since the molecules move out as the substrate comes in, the binding and removal of solvent molecules must be an intimate part of the reaction mechanism, one that can best be investigated by molecular dynamics.

A number of amino acids (for example histidine 119 and lysine 41) appear to be rather more flexible. Their relative positions in the native structure of ribonuclease are maintained primarily by a network of bridging water molecules, stabilized by the net charges on the side chains of these amino acids. As a result the amino acids are comparatively free to be repositioned in the presence of a substrate or a reaction intermediate. In this way they presumably take part more efficiently in the catalysis. A complete analysis of the RNA-cleaving reaction catalyzed by ribonuclease will require much more work, but even the current results show the importance of supplementing X-ray-crystallographic data with molecular-dynamics simulations.

The methodology and the results we have outlined make the future of molecular-dynamics simulations easy to predict. A wide range of biological problems involving proteins (not to mention nucleic acids and the lipid molecules in cell membranes) are ready for study. In coming years investigators should be able to learn how to calculate the rates of enzymatic reactions, the binding of small molecules to large ones and many other biologically important processes. The role of flexibility and fluctuations in the function of macromolecules will be understood in much greater detail. It should become possible to determine how particular solvent conditions and amino acid sequences produce certain patterns of protein fluctuations. Indeed, as the predictive abilities increase, the study of protein dynamics will have application to practical problems in endeavors such as genetic engineering or the effort to modify an enzyme to better serve an industrial process.






**BOUNDARY SIMULATION METHOD eliminates the need to simulate an entire enzyme in a "box" of water; instead one simulates the active site of the enzyme and introduces a boundary region that confines the active-site water molecules and allows an exchange of energy across the boundary between the active site and its surroundings. This computer drawing, prepared by Axel Brünger of Harvard, shows the active site of the enzyme ribonuclease and the essential water molecules (*orange and white*) at the active site. Shadows on the active site and on the background suggest the remaining part of the enzyme, not treated explicitly in the simulation. Water molecules observed in the simulation (*colored lines in bottom drawing*) form a network of hydrogen bonds (*colored dotted lines*) connecting charged side chains in the active site. In each bond a hydrogen atom is shared between a water molecule and an amino acid. Such bonds help to maintain the shape of the site.**

# The Quantized Hall Effect

*This variation on a classical phenomenon makes it possible, even in an irregular sample, to measure fundamental constants with an accuracy rivaling that of the most precise measurements yet made*

by Bertrand I. Halperin

Few areas of modern science have been studied as intensively as the physics of semiconducting materials. Semiconductors are key components of the electronic devices that are the hallmark of our era. In developing these devices, investigators have learned how to grow nearly perfect semiconductor crystals; how to modify the electronic properties of such crystals by the addition of a few parts per million of foreign atoms, and how to understand so well the properties of the modified crystals that they can incorporate them into circuit elements whose thickness is measured in tens of atomic layers and whose area is so small that a million individual elements can fit on a chip measuring one square centimeter.

It might seem unlikely, then, that in the 1980's semiconductor physicists should have been caught by surprise by the experimental discovery of a dramatic new effect, so little anticipated that many months would pass before experts in the field could develop a satisfactory explanation. In fact, there have been two such surprises.

In 1980 Klaus von Klitzing, working at the High Magnetic Field Laboratory of the Max Planck Institute in Grenoble, made a discovery that was to earn him the 1985 Nobel prize in physics: he observed a phenomenon that is now called the integral quantized Hall effect. Two years later, in an experiment at the National Magnet Laboratory in Cambridge, Mass., Daniel C. Tsui, Horst L. Störmer and Arthur C. Gossard of AT&T Bell Laboratories observed a second remarkable phenomenon, now known as the fractional quantized Hall effect. As their names imply, these effects have much in common, but they require significantly different explanations.

The quantized Hall effect, integral or fractional, is actually a special case of a more general phenomenon known as the Hall effect. To observe the Hall effect, a planar sample of conductive material is placed in a magnetic field perpendicular to the sample's surface. Then, by means of electrodes at each end of the sample, a small electric current is made to pass from one end to the other. (An electric current is simply the motion of charged particles; in the case of most solids the current-carrying particles are electrons.)

Two measurements are then made. To make one measurement, contacts near both ends of the sample are connected to a voltmeter, which measures the voltage drop from one end of the sample to the other, that is, in the direction parallel to the flow of current. (This is essentially a measure of the difference between the energy of the electrons entering the sample and the energy of those that have passed through it.) The other measurement is made by a voltmeter attached to electrodes positioned on opposite edges of the sample in such a way that a line drawn between the electrodes would be perpendicular to the direction of the current. This voltmeter measures the voltage difference across the sample, in the direction perpendicular to the current.

If there were no magnetic field, the voltage difference perpendicular to the current would be equal to zero. As the American physicist Edwin H. Hall first noticed in 1879, when there is a magnetic field perpendicular to the sample, a voltage difference occurs that is in most cases directly proportional to the strength of the magnetic field. The voltage drop parallel to the current is usually about the same whether or not there is a magnetic field.

The voltage drop parallel to the current and the voltage drop perpendicular to the current are both directly proportional to the amount of current flowing in the sample. It is therefore convenient to divide each of these voltages by the amount of current flowing and thus obtain quantities that are independent of the current flow. The ratio of the voltage drop along the sample to the current is just the ordinary electrical resistance of the sample. The ratio of the voltage drop across the sample to the current flowing along it is called the Hall resistance.

In his 1980 experiment von Klitzing was studying the Hall effect in certain devices in which the electrons free to carry current are confined within a thin layer. He found that when such devices are cooled to within about a degree of absolute zero and placed in extremely strong magnetic fields, the behavior of the ordinary resistance and the Hall resistance was dramatically different from the behavior described by Hall. He had found the integral quantized Hall effect.

The most striking difference between the integral quantized Hall effect and the normal Hall effect is that the Hall resistance, instead of increasing steadily and linearly as the strength of the magnetic field is increased, exhibits a series of plateaus. That is, there are intervals in which the Hall resistance appears not to vary at all when the strength of the magnetic field is varied. (Between such plateaus, the Hall resistance does rise smoothly with increasing magnetic field.) In addition, in the same intervals of magnetic field strength where the Hall resistance exhibits plateaus, the voltage drop parallel to the current seems to disappear completely. That is, there is no electrical resistance in the sample and current flows from one end to the other without dissipating any energy.

The vanishing electrical resistance and the plateaus in the Hall resistance are remarkable phenomena in themselves. The effect seems even more astonishing when one examines the values at which the plateaus in the Hall resistance appear. On each plateau the value of the Hall resistance satisfies a remarkably simple condition: the reciprocal of the Hall resistance is equal to an integer multiplied by the square of the charge on the electron and divided by Planck's constant, the fundamental constant of quantum mechanics. Each plateau is characterized by a different integer.

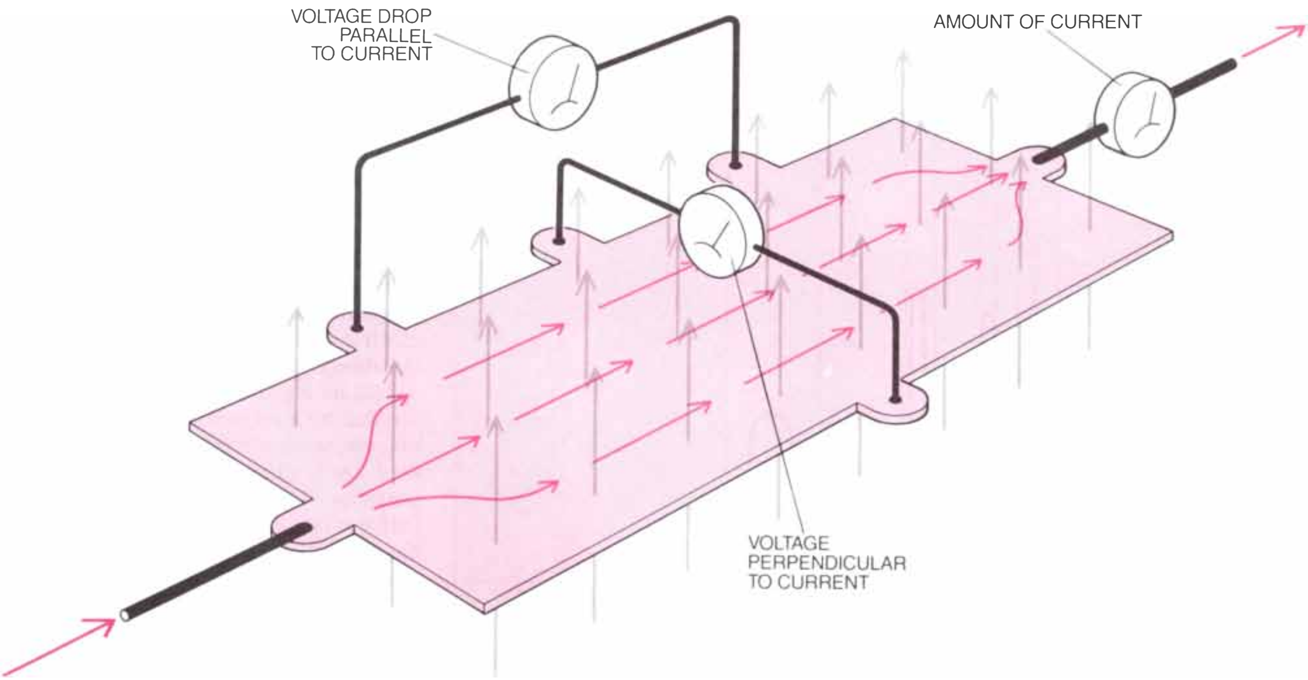
The fractional quantized Hall effect found by Tsui, Störmer and Gossard occurs only in devices that are exceptionally free from defects that interfere with the motion of electrons. In the fractional effect there are additional plateaus where the reciprocal of the Hall resistance is equal to a simple fraction (rather than an integer) multiplied by the square of the charge on the electron and divided by Planck's constant. In the original experiments the observed fractions were 1/3 and 2/3. More recent experiments have found plateaus at levels that correspond to such other fractions as 4/3, 5/3, 2/5, 3/5 and 3/7, but no plateaus have been found at levels corresponding to fractions with even denominators.

What is most amazing about the quantized Hall effect is its degree of precision. Measured values of the Hall resistance at various integer plateaus satisfy the simple relation I have described with an accuracy of about one part in six million—precisely the accuracy with which the values of the fundamental constants themselves are known. That the Hall resistance in the integral effect is precisely quantized (in other words, that individual values of the Hall resistance are equal to the same quantity multiplied by different integers) has been examined by determining the ratio of the Hall resistances at different plateaus and ascertaining that they are in the ratio of integers. The experiments that have determined such ratios are accurate to within one part in 30 million.

The precision of the quantized Hall resistance is particularly surprising because under normal conditions the Hall resistance of a sample, like the ordinary electrical resistance, is notoriously sensitive to such details as the sample's geometry, its processing history, the amount of impurity it contains and its temperature at the time of the experiment. The quantized Hall effect, on the other hand, is the same for samples with different geometries and processing histories, and even for those made of a variety of materials. It is almost without precedent for any property of a solid to be reproducible from one sample to another at the levels of accuracy attained by the quantized Hall effect.

The integral and fractional quantized Hall effects are found only in semiconductor devices in which the electrons that are free to carry current are confined within a thin layer of the semiconductor. There are two common devices that constrain electrons in this way. In the original device studied by von Klitzing, which is called a field-effect transistor, a metal contact is separated from a silicon substrate by a thin layer of insulating material. A voltage applied to the metal contact can attract electrons to a thin layer just under the surface of the silicon. (The high-quality samples employed by von Klitzing were constructed by Gerhardt Dorda of the Siemens Research Laboratory in Munich and Michael Pepper of the University of Cambridge.)

The device in which the fractional quantized Hall effect was first seen, which is called a heterojunction, is an interface of crystals made out of two different semiconducting materials. In a heterojunction, electrons from



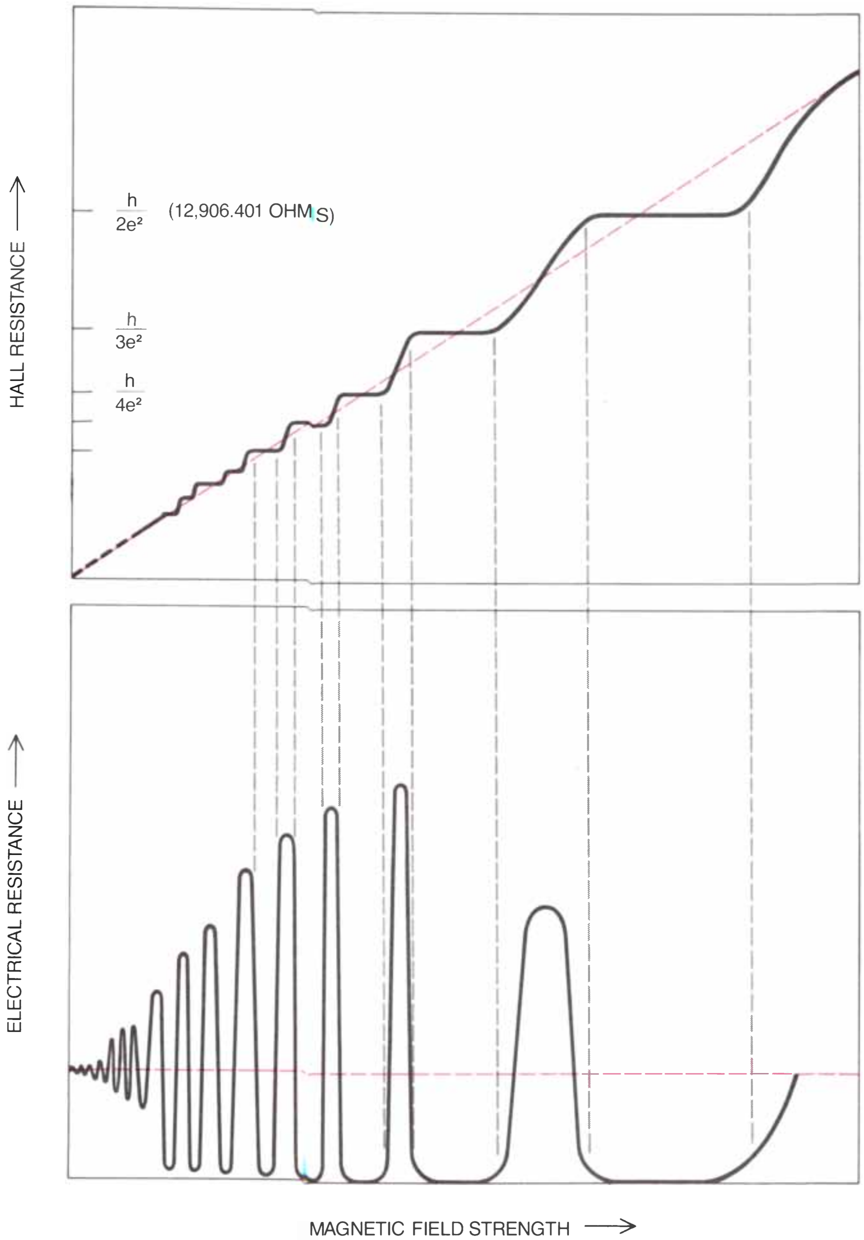

**APPARATUS for measuring the Hall effect consists of a flat plane of conductive material and devices for measuring voltages and currents. A measured amount of current (*horizontal arrows*) is made to flow from one end of the sample to the other and a magnetic field (*vertical arrows*) is applied perpendicular to the sample. Two voltages are measured: the voltage drop from one end of the sample to the other (which is essentially the difference in energy between the electrons entering the sheet and those leaving it) and the voltage perpendicular to the current, which is called the Hall voltage. In the classical Hall effect the Hall resistance (the Hall voltage divided by the amount of current flowing) is directly proportional to the strength of the magnetic field: it increases steadily as the magnetic field strength is increased. In the quantized Hall effect, which is seen when the electrons that carry current are confined to a thin layer of a semiconductor device at low temperatures and in extremely strong magnetic fields, the Hall resistance shows plateaus: there are intervals of magnetic field strength in which the Hall resistance remains constant as the magnetic field is varied. Moreover, in those intervals the sample's electrical resistance (the voltage drop parallel to the current divided by the amount of current) disappears completely, indicating that current flows through the sample without dissipating any energy. The values of the Hall resistance on the different plateaus are exactly the same regardless (within broad limits) of the geometry or even the composition of the device.**

one semiconductor are attracted to more energetically favorable locations in the other semiconductor. The positive charge thereby created in the "donor" semiconductor provides a force attracting the electrons back, however, and they become trapped in a thin layer at the interface of the two crystals.

In both the field-effect transistor and the heterojunction the current-carrying electrons are free to move anywhere within the plane to which they are confined. Motion perpendicular to this plane, however, is almost completely impossible, and so for all practical purposes the layer of electrons may be regarded as two-dimensional. In experiments on the quantized Hall effect this thin layer of electrons plays the role the flat sheet of conductor plays in normal Hall experiments.

Although the properties of two-dimensional layers of electrons have been studied thoroughly, the integral quantized Hall effect was a great surprise when it was discovered. Some aspects of the effect had actually been predicted as an incidental result in a paper on a related subject written in 1975 by Tsuneya Ando, Yukio Matsumoto and Yasutada Uemura of the University of Tokyo, but the most important aspect, the extreme precision of the effect, had not been expected because the predictions were based on a type of analysis that is in general only approximately correct for the electrons in real materials. For the fractional quantized Hall effect there was not even an approximate explanation available at the time of its discovery.

Although there are now several convincing explanations of the exactness of the quantized Hall effect, it is not possible to present in this article a complete explanation of the fractional or the integral quantized Hall effect. Any such explanation would necessarily be written in the highly mathematical language of quantum mechanics, which is the basis for all modern descriptions of phenomena on the atomic scale. It is possible, however, to give a partial explanation that incorporates many of the essential ideas and that will at least make plausible the experimental results.




**QUANTIZED HALL EFFECT appears as plateaus in the Hall resistance of a sample (*top*), which coincide with the disappearance of the sample's electrical resistance (*bottom*). On the plateaus the Hall resistance does not change when the magnetic field strength is varied. In contrast, in classical experiments (*color*) the Hall resistance increases linearly with increasing magnetic field strength and the resistance is constant. At each plateau the value of the Hall resistance is precisely equal to Planck's constant (the fundamental constant of quantum mechanics, designated by the letter *h*) divided by an integer multiple of the square of the charge on the electron. The quantized Hall effect thus makes possible precise measurements of fundamental constants and extremely accurate calibration of instruments.**

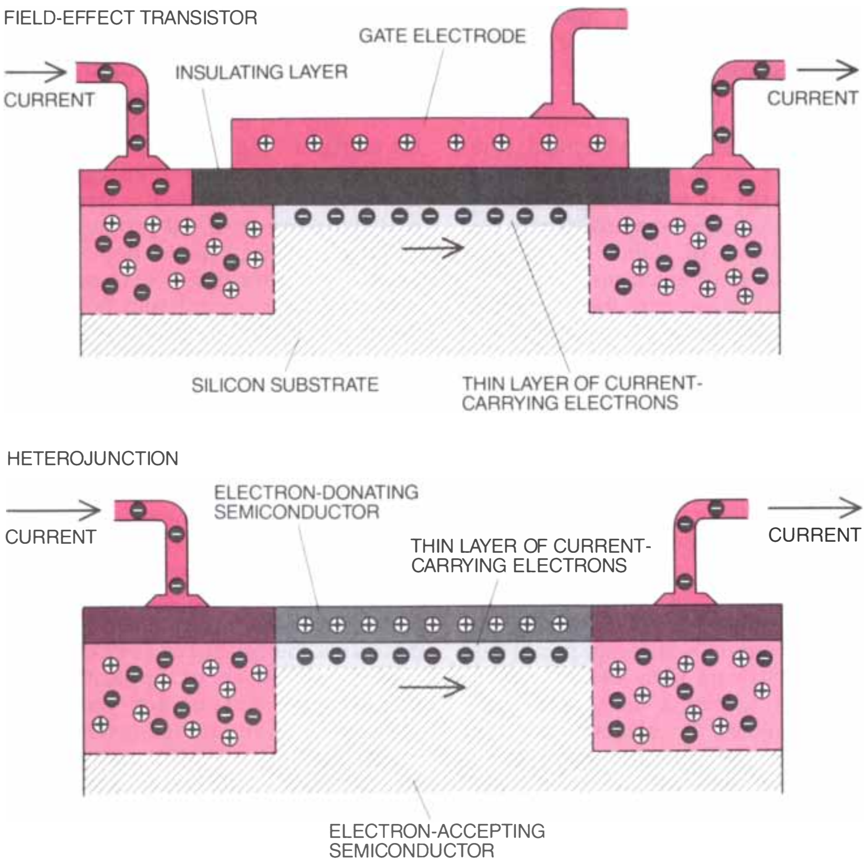
Essential to any exposition of the Hall effect are such basic classical concepts as electric and magnetic fields. An electric field is a field that exerts a force on charged objects, much as a gravitational field exerts a force on massive objects. When a marble rolls down a hillside, it tends to roll in whatever direction the hill is steepest; similarly, a charged particle in an electric field generally feels a force in the direction of the field lines.

Just as a gravitational field may be created by a concentration of massive objects, so an electric field may be created by a concentration of charged particles. The analogy can be carried further: just as an object high on a hill has a greater "gravitational potential energy" than an object in a valley, so a charged particle may have a greater "electrical potential energy" than a similar particle elsewhere in an electric field. The difference between the electrical potential energy of similar charged particles at two different points in space is called the voltage difference between those points.

Unlike an electric field, a magnetic field exerts a force only on moving charged particles. The magnitude of the force exerted on a particle is directly proportional to the particle's velocity and the strength of the magnetic field. The direction of the force is perpendicular to both the particle's direction of motion and the direction of the magnetic field. A moving charged particle in a magnetic field therefore moves in a circle: at each instant the magnetic field exerts a force perpendicular to the particle's motion, and so at each instant the particle's path curves. The radius of the circular orbit is inversely proportional to the strength of the magnetic field, so that stronger magnetic fields hold particles in smaller orbits.

Suppose a moving charged particle is placed simultaneously in both an electric field and a magnetic field. In this case the particle follows a cycloidal path: it moves in a circle around a so-called guiding center, which itself moves in a straight line [*see illustration on next page*]. It turns out that the guiding center of the particle's motion moves in a line perpendicular to both the electric field and the magnetic field. In other words, it moves neither up nor down the "hillside" of the electrical potential energy. Rather, on the average it tends to move horizontally along the hillside, in the direction perpendicular to that of the slope.

The speed with which the guiding center moves, called the drift velocity, is inversely proportional to the strength of the magnetic field—when the magnetic field is stronger, the particle's motion is more circular and less linear—and directly proportional to the strength of the electric field; thus the guiding center moves faster in regions where the "incline" is steepest, even though it moves along the slope, not down it.




**SEMICONDUCTOR DEVICES in which the quantized Hall effect is observed hold current-carrying electrons within a thin layer of semiconducting crystal. In a field-effect transistor (*top*) electrons are attracted to the surface of a silicon crystal by a positive charge applied to a metal contact called the gate electrode. Insulating material prevents the electrons from entering the electrode, and they remain in a thin layer near the crystal's surface. In a heterojunction (*bottom*), an interface of semiconductors with differing electronic properties, electrons from one semiconductor are attracted to lower-energy states in the other. The positive ions left behind attract the electrons back, however, trapping them in a thin layer.**

With these basic principles it is possible to understand the classical Hall effect. Suppose a current of electrons is flowing from left to right on a finite sheet of metal that lies on a plane parallel to this page. Now suppose a magnetic field is applied perpendicular to the metal sheet. At first the electrons will be pushed in a direction perpendicular both to their original direction and to that of the magnetic field: they will move toward the top or bottom edge of the page. If electrons cannot leave the metal except through the contacts at the left and right ends, they will accumulate at one edge (top or bottom) of the sheet. This accumulation of charged particles produces an electric field that points from one edge of the metal to the other, and so one edge will be at a higher voltage than the other.

The electric field, in combination with the magnetic field, impels electrons to move in cycloidal paths in a direction perpendicular to both fields: from left to right. The voltage difference between the two edges of the metal is therefore perpendicular to the direction in which current flows.

If the strength of the magnetic field is increased, then for a given amount of current to flow (that is, for electrons to continue to flow from left to right at the same rate) the strength of the electric field—and the voltage difference between the metal's edges—must also increase. The reason is that the "drift velocity" with which the guiding center of each electron's cycloidal path moves from left to right is inversely proportional to the strength of the magnetic field and directly proportional to the strength of the electric field. Consequently when the strength of the magnetic field is increased, more electrons are deflected toward one edge of the metal, creating a stronger electric field and a greater voltage difference between the edges. The voltage difference, which increases linearly with an increase in the strength of the magnetic field, is the classical Hall effect.

This kind of classical analysis will not explain the quantized Hall effect, which is inherently a quantum-mechanical phenomenon. It is possible to understand some aspects of the quantized Hall effect, however, by means of what is called semiclassical analysis, a mixture of classical and quantum-mechanical ideas that avoids the mathematical complexity of a full quantum-mechanical description. Although a semiclassical analysis is not fully reliable, it is nonetheless effective as a guide that suggests the general form of the quantum-mechanical answer, and in fact it is often relied on for this purpose by experts in the field.

In the classical case an electron in a uniform magnetic field follows a closed circular orbit. In quantum mechanics the energy of a particle in such a closed orbit can have only certain discrete values, just as an electron in an atom can occupy only certain energy levels. The energy levels of the two-dimensional motion of a charged particle such as that of an electron in a uniform magnetic field are known as Landau levels. There is no discrete set of energy levels for thick, "three-dimensional" systems in a uniform magnetic field, because the energy of the particle's motion parallel to the magnetic field, unlike that of its motion in the plane perpendicular to the field, is not quantized. It is therefore significant, for the sake of this analysis, that the actual physical system under consideration consists of a thin, "two-dimensional" layer of electrons.

In the classical case there are many places within the plane where the electrons' identical circular orbits could be centered without overlapping one another. Likewise, in the semiclassical case there are a number of states with the same energy that are independent of one another. (For each state there is a distinct "wave function," a mathematical description of the electron's probable position.) Specifically, the number of independent states per unit of area is directly proportional to the strength of the magnetic field. This is at least plausible, because in the classical case a stronger magnetic field confines the electrons to tighter orbits, and so more orbits can fit in a given area without overlapping. It turns out that the constant of proportionality between the number of independent states per unit of area and the strength of the magnetic field depends only on Planck's constant and the charge of the electron.

So far I have ignored an aspect of the actual experimental system that is very important in all properties involving electrical conduction: the influence of impurities or defects in the semiconductor crystal that holds the thin plane of conduction electrons. Impurities play a particularly important role in the ordinary electrical resistance of metals and semiconductors. Much of the energy dissipation that characterizes resistance occurs when electrons are scattered by collisions with impurity atoms or defects in the crystal lattice. Paradoxically, the presence of impurities is what leads to the disappearance of electrical resistance and the plateaus in Hall resistance that constitute the quantized Hall effect.

In the presence of impurities the many independent quantum states that make up a given Landau level are no longer precisely equal in energy. In a semiclassical explanation one might say, for example, that in certain of the quantum states the electron is slightly more likely to be found near an impurity atom that has, say, an excess of positive charge. (This would be the case if the impurity atom were of an element that tends to give up an electron when it is buried in a semiconductor lattice.) Such states would be slightly stabler than others in the same Landau level and would have a slightly lower energy. The single energy level that makes up a Landau level in a pure crystal is thus spread out, in the presence of impurities, into a band made up of many distinct energy levels.

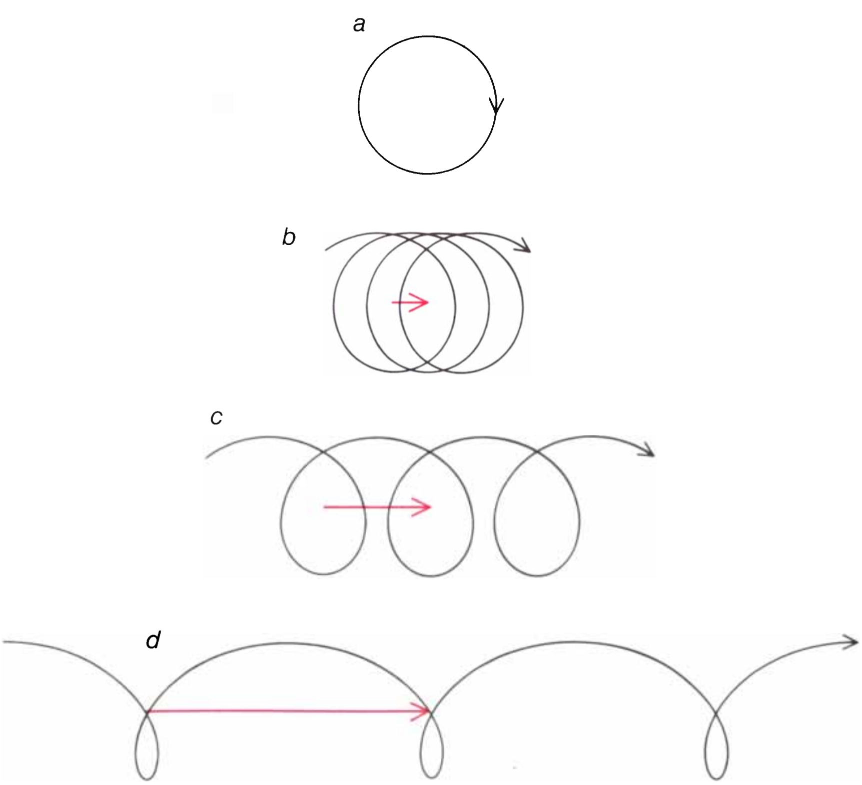
The various quantum states in each energy band can be divided into three general classes. The states near the bottom of each band, that is, those of lowest energy, are each localized in some small region of the sample. An electron in such a state would never leave that region. These low-energy localized states occur at "valleys" in the electrical potential energy, for example in the regions around impurity atoms that have an excess of positive charge. Similarly, near the top of the band are high-energy localized states. These are localized near "peaks" in the electric potential, perhaps in the regions around impurity atoms that have acquired electrons and so bear an excess of negative charge.

Near the center of each energy band are the so-called extended states, each of which is spread out over a large region of space. The distinction between localized and extended states is important: because electrons in localized states cannot move very far within the sample, they cannot carry current [*see illustration on pages 58 and 59*].

Taking into consideration one more aspect of the actual experimental system, the extremely low temperatures at which the experiments are conducted, will now make it possible to give some explanation of the integral quantized Hall effect.

At very low temperatures the electrons in the sample take on the configuration that has the lowest possible energy. According to the Pauli exclusion principle, one of the foundations of quantum mechanics, no two electrons can occupy the same quantum-mechanical state. When the system is at its lowest possible energy, then, each available quantum state below a certain energy level contains exactly one electron and each state above that level contains no electrons. The energy of the highest filled energy level is known as the Fermi energy.

When there is a voltage difference between the two edges of the sample, it is not actually possible to define a sin-




**CYCLOIDS describe the motion of an electron in combined electric and magnetic fields. In the absence of an electric field an electron in a magnetic field moves in a circle in the plane perpendicular to the field (*a*). When an electric field is applied, the electron follows a cycloid: it moves in a circle around a guiding center that itself moves in a straight line. The line followed by the guiding center is perpendicular to both the electric field and the magnetic field. The speed with which the guiding center travels is inversely proportional to the strength of the magnetic field and directly proportional to the strength of the electric field; when the electric field is stronger or the magnetic field is weaker (*b–d*), the electron's motion is more linear and the guiding center moves with greater "drift velocity" (*colored arrows*).**

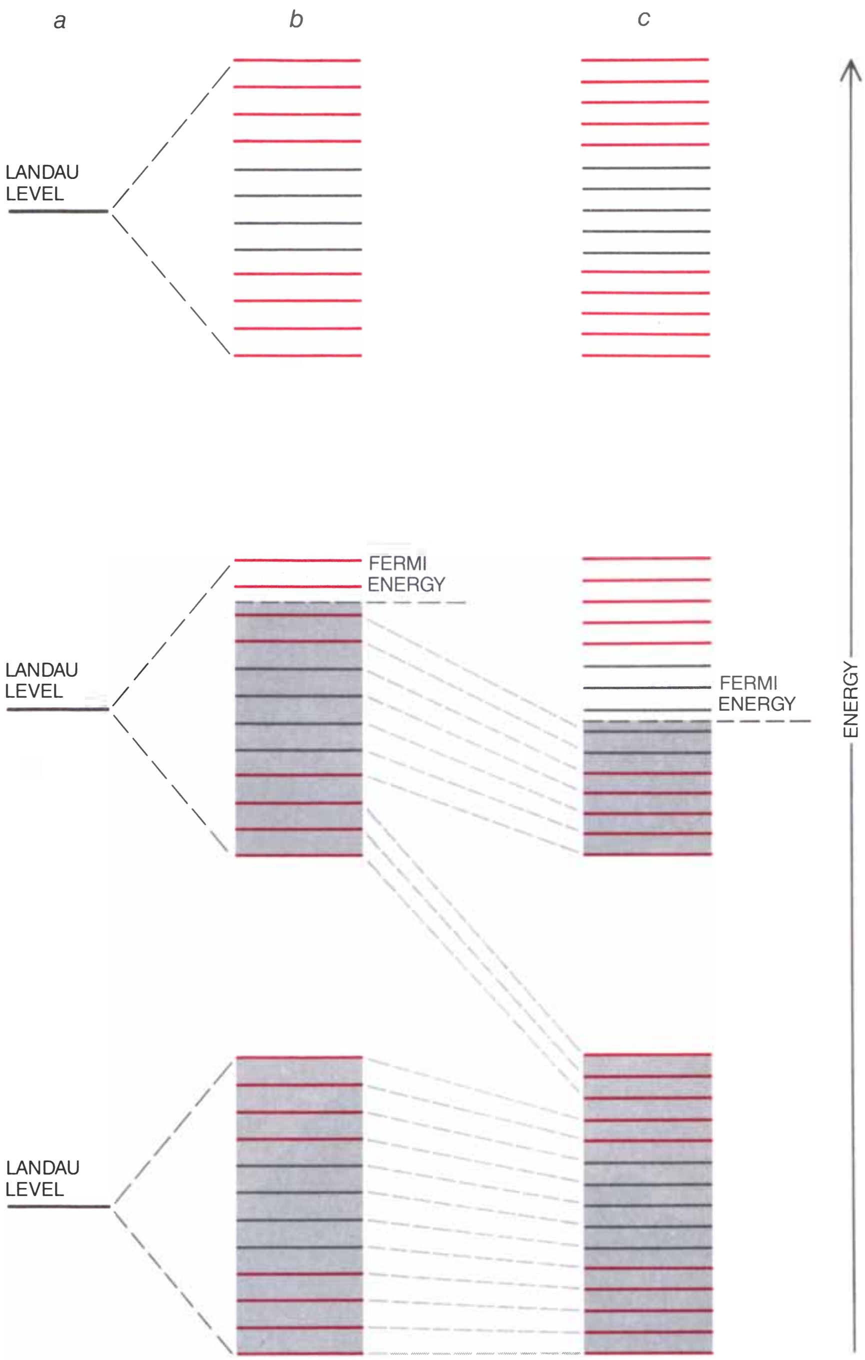
gle Fermi energy for the entire plane of conduction electrons; the Fermi energy varies slightly from point to point. In each region of space the states with energies below the local Fermi level are occupied and states above it are empty. The voltage measured by a voltmeter connected to two points on opposite edges of the sample is actually a measure of the difference between the local Fermi energies of the two points; the difference between the local Fermi energies represents the difference between the energies of the highest-energy electrons at each point.

Suppose a current is flowing along a sample in a perpendicular magnetic field and the Fermi level of the sample's electrons is in the sub-band of localized states near the top of some Landau band. In this case all the extended states and the low-energy localized states of the Landau band will be completely filled, and some of the high-energy localized states will be occupied as well. Now suppose that the strength of the magnetic field is gradually increased and that at the same time the current is continuously adjusted in such a way that the Hall voltage between the sample's two edges remains constant.

Because the number of independent quantum states per unit of area is directly proportional to the applied magnetic field, as the magnetic field increases, the number of independent quantum states in each Landau level increases proportionately: in every region of space within the sample, additional quantum states that have roughly the same energy as neighboring states become available.

Many of the newly available states will be below the local Fermi level, and so electrons from higher-energy occupied states will drop down to fill them; these electrons will in general come from the high-energy localized states that are near the Fermi level. As these states are vacated the Fermi level—the energy of the highest occupied state—descends to a lower position within the Landau band [*see illustration at right*]. As long as the Fermi level remains in the sub-band of high-energy localized states, all the extended states within the Landau band remain fully occupied.

Because an electron trapped in a localized state cannot move through the sample, the changing fraction of localized states that are filled has no effect on the sample's large-scale electrical properties. The amount of current flowing in the sample therefore remains constant as long as the sub-band of extended states is completely filled: although the increased magnetic field slows the forward motion of any cur-




**ENERGY LEVELS** available to an electron confined in a two-dimensional layer in a perpendicular magnetic field determine the layer's large-scale electrical properties. In a layer in an ideal crystal (*a*) an electron can exist only in states that have certain energy levels, called Landau levels. In real crystals, which always contain impurities, each Landau level is spread into a band of energy levels (*b*); states in the same Landau level, which had once had equal energies, come to have slightly different energies because in some states the electron is closer to an impurity atom. The energy levels are of two types: localized states (*color*), in which the electron is bound in the vicinity of an impurity atom or defect, and extended states, in which the electron can roam over a large area of the crystal. Localized states can have high or low energies, depending on the charge of the impurity producing them. Only electrons in extended states can carry current, and so the crystal's electrical properties are determined by the electrons in extended states. Because of the Pauli exclusion principle, each state may contain no more than one electron. At very low temperatures every state below a certain energy, the Fermi energy, is occupied (*shaded region*) and every state above it is empty. When the strength of the magnetic field is increased, there are more states in each Landau band, and electrons drop (*broken lines*) into states with lower energies (*c*). The Fermi energy thereby drops to a lower position within the Landau band or to a lower Landau band.

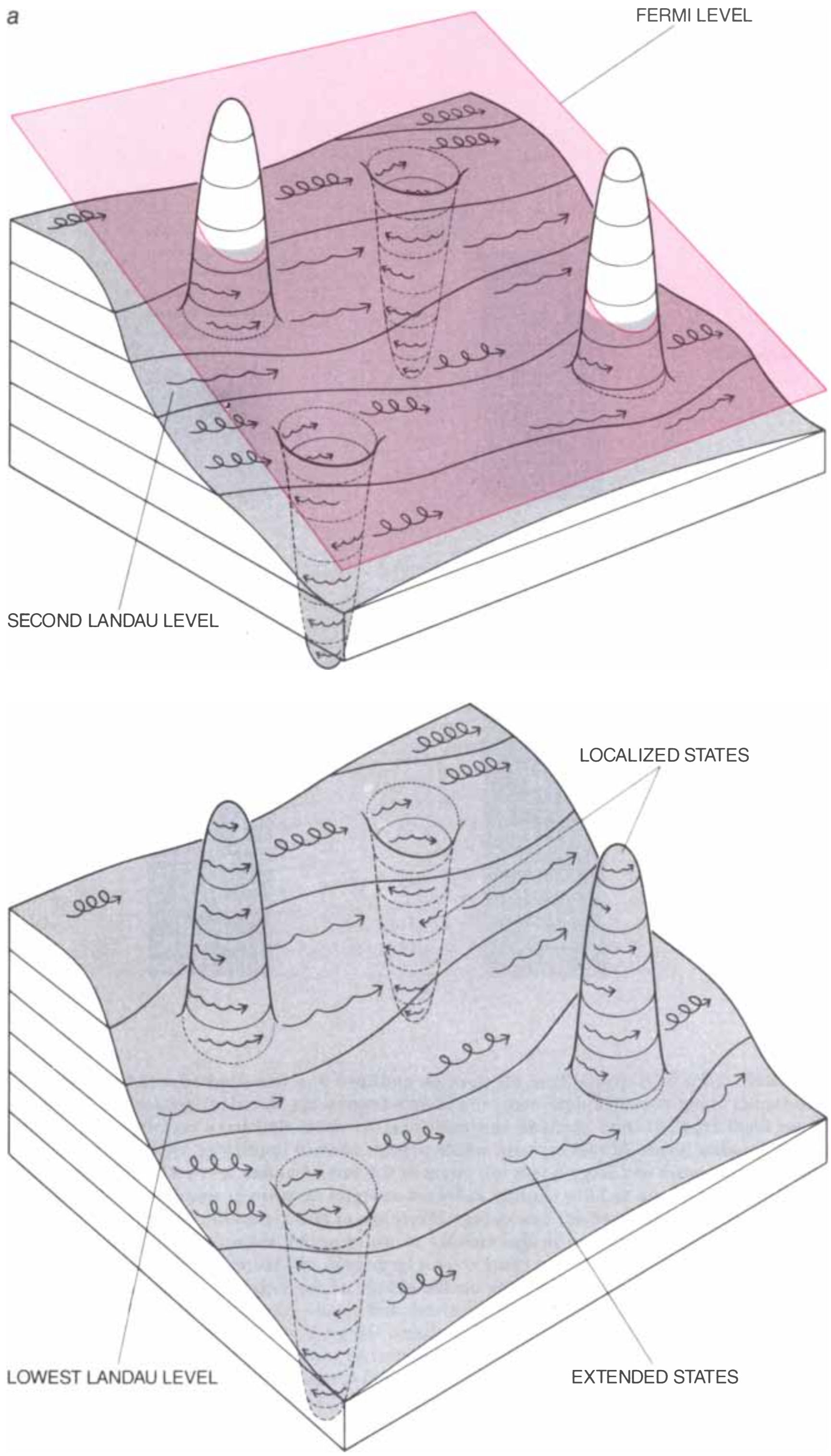
rent-carrying electrons, this effect is precisely canceled by the increase, due to the newly created extended states, in the number of electrons available to carry current.

Since the Hall voltage is being held constant, the fact that the current does not change as the magnetic field is varied implies that the Hall resistance (the Hall voltage divided by the amount of current flowing) also remains constant. (An equivalent experiment, in which the current is externally held constant and the Hall voltage is measured, can also be performed; in this experiment the Hall voltage, and hence the Hall resistance, remains constant whenever the Fermi level is in the sub-band of localized states.) Whenever the Fermi level is in the sub-band of localized states, then, the Hall resistance remains the same even when the magnetic field is varied. This is the plateau in Hall resistance that is characteristic of the quantized Hall effect.

Eventually, as the strength of the magnetic field is increased, the supply of electrons in the high-energy localized states will be exhausted and the Fermi level must drop into the sub-band of extended states. As the Fermi level descends through the sub-band of extended states, some of them are vacated. Because the current-carrying sub-band is then only partially occupied, the amount of current flowing decreases and the Hall resistance therefore increases. The Hall resistance continues to increase when the magnetic field is increased, as long as the Fermi level remains in the sub-band of extended states.

If the magnetic field is increased further, eventually the extended states within the Landau band will all be emptied, and the Fermi level will once again enter a sub-band of localized states: the low-energy localized states at the bottom of the Landau band. If there is at least one full Landau band below the Fermi level, the extended states in that band will be able to carry a current and the quantized Hall effect will once more be observed. Because the extended states of one Landau band have been completely emptied, however, the number of sub-bands of occupied states has been reduced by one; the electric current is therefore reduced in precise proportion to the number of occupied sub-bands of extended states, and the Hall resistance is accordingly larger than it was on the previous plateau.

In this model it is easy to understand why the ratio of Hall resistances at any two plateaus should equal a ratio of integers. The reason is that for any given Hall voltage the current is




**GEOMETRIC REPRESENTATION of the energy of electrons confined to a two-dimensional layer in a strong magnetic field illustrates the difference between localized and extended states and explains the quantized Hall effect. The vertical axis represents the energy, within a small section of a sample, of electrons in two Landau bands. Peaks and valleys in the electric potential are due to impurities spaced randomly within the sample. Electrons move in cycloidal paths parallel to contour lines (lines of constant electric potential); hence closed contours around peaks and valleys indicate localized states, in which electrons orbit within a small region of space, and open contours indicate extended states, which may stretch from one end of the sample to the other. At low temperatures electron states whose energy lies below the local Fermi level (*colored plane*) are filled (*shad-***

directly proportional to the number of occupied sub-bands of extended states, and on each plateau an integral number of such sub-bands is filled.

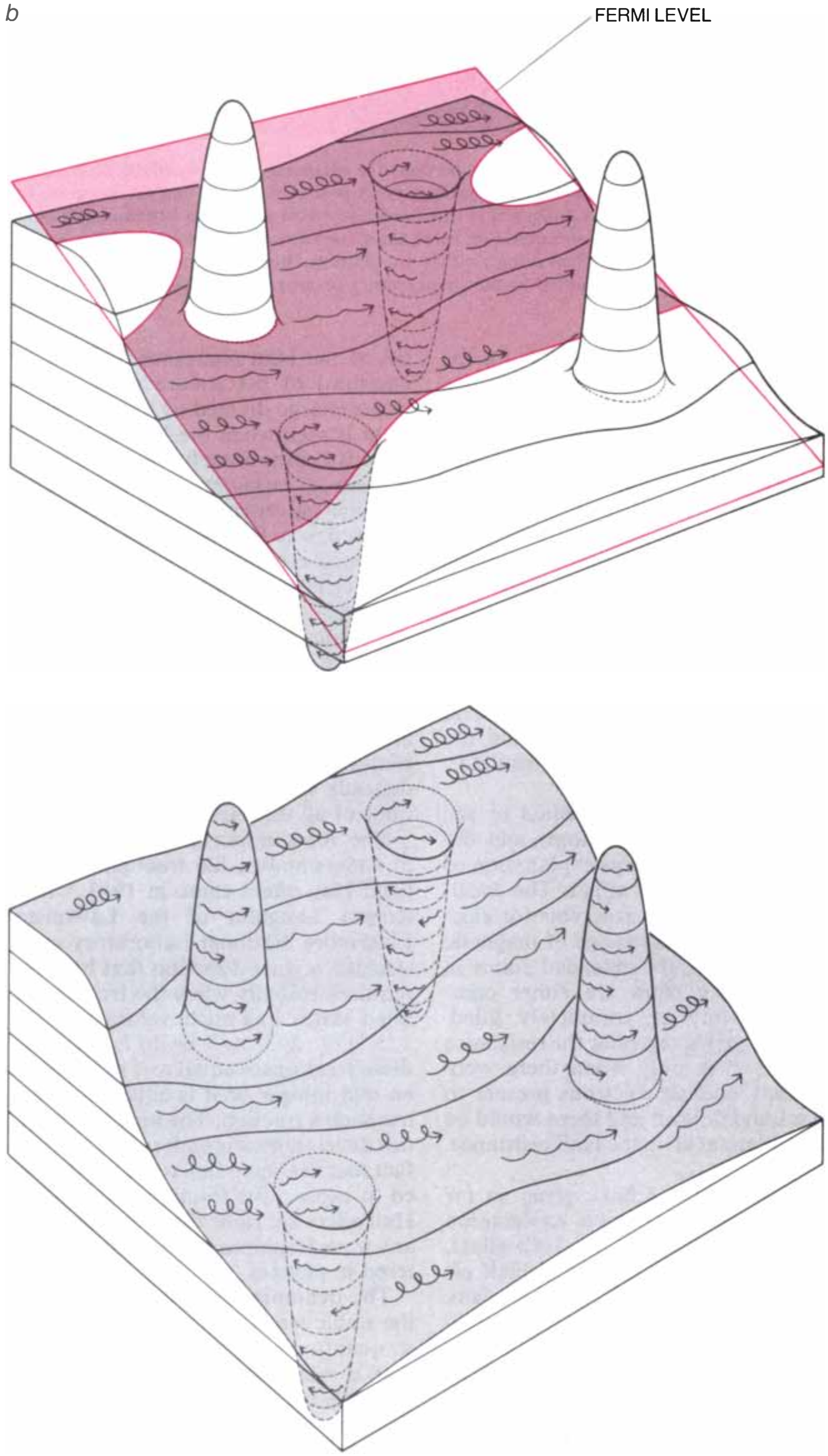
If the magnetic field is increased still further, the Fermi level will move down through the regions of localized states at the bottom of one Landau band and into the high-energy localized states at the top of the next Landau band. The Hall resistance will remain constant at its new plateau value until the Fermi level reaches the region of extended states in the middle of this next Landau band.

The second striking feature of the quantized Hall effect is that current flows through the sample with no resistance, that is, without dissipating any energy. This feature can be explained by the principle of conservation of energy.

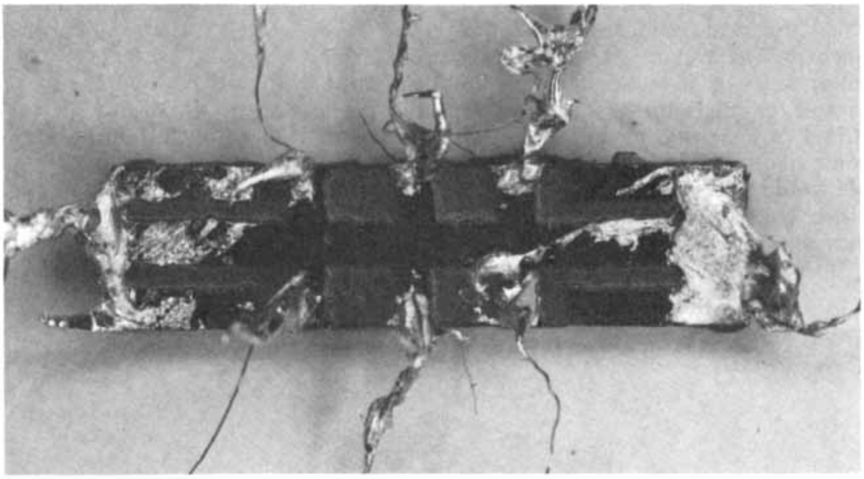
To dissipate power an electron must make a transition from one energy state to a state of lower energy, dissipating the excess energy as vibrational energy, or heat, within the crystal lattice. Because two electrons cannot occupy the same quantum-mechanical state, the new state into which the electron falls must be one that was previously empty. If the Fermi level is in the sub-band of energies corresponding to the extended states, an electron can find itself in an extended state that is below the local Fermi energy in one region of space but that extends into a region where its energy is above the local Fermi energy. In that region there may be states with energies between the energy of the electron's state and the local Fermi energy. The electron would then be able to drop into one of those states.

In particular it could happen that an electron enters the sample in a region where the local Fermi energy is high and exits from the sample in a region where the local Fermi energy is lower. Such an electron might drop into a lower level as it passes through the sample, exiting with less energy than it had on entering, having dissipated some of its energy as heat. The sample would thus exhibit electrical resistance: electrons would lose energy as they passed through it.

If the Fermi level is within the energy levels corresponding to the localized states, however, the sample cannot exhibit electrical resistance. The electrons with the highest energy—those whose energy is near the Fermi level—will all be localized in very small regions of space. In those regions all the unoccupied states must have energies higher than the local Fermi energy. There may be empty states of lower energy in some distant part of




**_ed region_) and others are empty. The surfaces of electron energy and the local Fermi level are tilted from the horizontal because of a voltage difference (the Hall voltage) between the two edges of the sample. When the Fermi level is in a region of localized states (_a_), the quantized Hall effect occurs: the Hall resistance (the Hall voltage divided by the current flow) is not affected by small changes in such parameters as the strength of the magnetic field or the number of electrons in the layer. For example, a small change in the Fermi level would not change the Hall resistance because it would not affect electrons in the extended, current-carrying states. When the Fermi level is in a region of extended states (_b_), on the other hand, a small change in its height does indeed affect the current flow by emptying or filling some of the extended states, and the quantized Hall effect does not occur.**



**EXPERIMENTAL DEVICE (a heterojunction) for measuring the quantized Hall effect has irregular features and imprecise geometry, yet it produces results that are accurate to within about one part in six million (the same accuracy with which the fundamental constants of electromagnetism and quantum mechanics are known). The most astonishing aspect of the quantized Hall effect is that, within broad limits, the extreme accuracy of the experimental results is independent of the shape, purity or processing history of the sample.**

the sample, but it is impossible for the electron to make a transition between localized states that are far apart in space. The electrons therefore cannot drop into states of lower energy, and so they cannot dissipate energy.

This explanation of the absence of ordinary resistance presupposes one important condition: that there exists at least one Landau band, below the Fermi energy, that has occupied extended states. These states carry the current that passes without dissipation. If the Fermi level descends until all the states below it are localized, no current will be able to flow and the sample will act as a perfect insulator, not a perfect conductor.

The localized states created by the presence of impurity atoms and defects are crucial to this explanation of the quantized Hall effect. The localized states act as a reservoir for electrons, so that, for a range of magnetic field strengths, the extended states in each Landau band are either completely empty or completely filled. Without such a reservoir the resistance would vanish only when there were precisely enough electrons present to fill a Landau level and there would be no plateaus at all in the Hall resistance.

The argument I have given so far represents a plausible explanation of the integral quantized Hall effect, but the fractional quantized Hall effect is much more difficult to explain. The integral effect is seen most readily when an integral number of Landau levels are filled. The fractional effect, on the other hand, is seen only when a Landau level is partially filled. For example, a plateau on which the reciprocal of the Hall resistance is equal to one-third of the square of the electron's charge divided by Planck's constant is seen when the lowest Landau level is approximately one-third full.

Even a completely quantum-mechanical treatment of the model I have adhered to so far—that of independent electrons moving under the influence of electric and magnetic fields—shows no special stability when a fraction of the available states are filled. To explain the fractional quantized Hall effect it is necessary to take into account the mutual interactions of electrons, and this means one must apply a quantum-mechanical wave function that depends simultaneously on the positions of all the particles in the system.

The most important breakthrough in understanding the fractional quantized Hall effect came in 1983, when Robert Laughlin of the Lawrence Livermore National Laboratory constructed a wave function that had the required stability when the fraction of filled states had such values as 1/3, 1/5, 1/7, 2/3, 4/5 or 6/7. Each of these fractions is equal to 1 divided by an odd integer or it is equal to 1 minus such a fraction. The importance of this development is underscored by the fact that the wave functions constructed to explain the fractional quantized Hall effect are radically different from any wave functions previously encountered in physics.

The denominator is odd in each of the stable fractions in part because of a quantum-mechanical requirement that is related to the Pauli exclusion principle but concerns a component of the wave function called its phase, for which there is no classical analogue. After Laughlin's original work, other theorists, including Laughlin and Duncan Haldane, who was then at the University of Southern California, and I were able to propose wave functions that are stable for other fractions having odd denominators, such as 2/5, 2/7 and 3/7.

An interesting theoretical prediction resulting from Laughlin's analysis is that if one adds an extra electron to a system in which the Landau level is already one-third full, the extra charge should appear in three separate places in the sample. Precisely one-third of the charge on the electron should appear in each place. Similar fractional charges should appear if one adds or removes electrons at any of the fractional quantized Hall plateaus. The fractional charges are called quasiparticles, and they are expected to behave very much like charged particles.

In the fractional quantized Hall effect it is the states of the fractionally charged quasiparticles, rather than those of electrons, that are localized by the hills and valleys in the potential that exist in the region of impurity atoms or defects. The localized states then act as the reservoir that maintains a constant Hall resistance for a finite range of magnetic field strengths.

It is rather unlikely that the quantized Hall effect will ever be employed in a device of great economic significance. The need for extremely strong magnetic fields and extremely low temperatures adds an expense that rules out most practical applications. Investigators will surely take advantage of the Hall plateaus in the laboratory, however, as precise standards of electrical resistance; in addition the quantized Hall effect will probably be incorporated into experiments establishing more precise values for the fundamental constants of quantum mechanics, electricity and magnetism. When they are combined with other measurements, these experiments may lead to more stringent tests of the fundamental theory of electromagnetic interactions, complementing tests of that theory carried out with high-energy accelerators.

The true importance of the quantized Hall effect does not lie in any of these applications, however, but rather in the new insight physicists have gained into the peculiar properties of systems of electrons in strong magnetic fields and into the hidden regularities implied by the mathematical laws of quantum mechanics. There is also value in the reminder that nature may well hold in store other surprising states of matter that none of us yet imagine.


DODGE LANCER
5/50 PROTECTION, STANDARD.

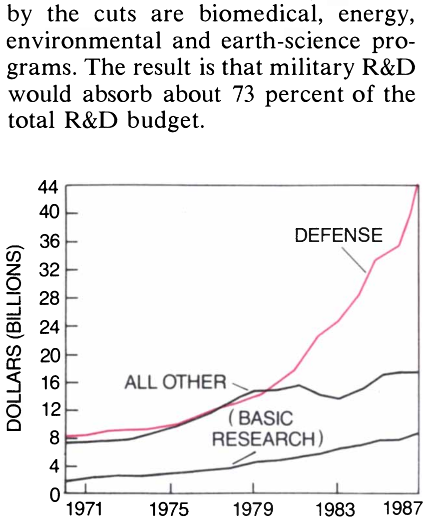
# SCIENCE AND THE CITIZEN

## *Coping and Hoping*

Facing the budget-cutting imperatives of the Gramm-Rudman Act and the steep tilt of the Reagan Administration's 1987 budget toward military research, those government agencies, universities and private laboratories that rely on Federal support for nonmilitary research and development are casting about for ways to cope with straitened circumstances. As they do so they hope attitudes in Congress toward the Administration's budget and certain trends in the economy may make their task easier.

Under the Gramm-Rudman Act, which calls for the Federal budget to be in balance by 1991, the deficit for fiscal 1987 must not exceed $144 billion. The current estimate for fiscal 1986 is $220.5 billion, minus $11.7 billion to be carved off by the first stroke of the Gramm-Rudman ax. To travel the rest of the way to $144 billion the Administration has proposed to terminate about 40 programs for 1987 and to make sharp cuts in many others.

A casual look at the budget suggests that science and technology will escape the ax. The 1987 budget proposes an increase of about 16 percent (from $54 billion in 1986 to $63 billion in 1987) in R&D programs. A look at the details, however, reveals a massive increase for military R&D, a significant increase for the physical sciences and engineering and cuts for virtually everything else except some programs in such fields as physics and computer science that are sheltered by the Administration's Strategic Defense Initiative umbrella. Most severely affected by the cuts are biomedical, energy, environmental and earth-science programs. The result is that military R&D would absorb about 73 percent of the total R&D budget.




**MILITARY PROGRAMS absorb most of the money proposed for the support of research and development in the 1987 budget.**

Investigators in vulnerable fields are hoping for help from fate. One hope is that the law itself may be declared unconstitutional. A three-judge panel of the U.S. District Court has already ruled to that effect, and the U.S. Supreme Court is expected to review the decision this spring. It is also possible Congress will revise the budget, shifting funds to nonmilitary programs.

Other developments can be discerned that could close the gap between Federal income and outgo, lessening the need to reduce the deficit by large spending cuts. One is the possibility that Congress will enact a tax increase in spite of the Administration's strong resistance. Another is that interest rates will go down; interest payments on the national debt now absorb 20 percent of the Federal budget. A third prospect is that the decline in oil prices will stimulate the economy, thereby increasing the Government's tax revenues. The decline will at least reduce Federal expenditures: the Department of Defense alone buys some 200 million barrels of oil per year.

What can the R&D community do if none of these things happens? Robert Rosenzweig, president of the Association of American Universities, pointed out that because of the decline in interest rates and other factors, the across-the-board cuts Gramm-Rudman will require for fiscal 1987 are now estimated at 8 percent, which is significantly lower than earlier estimates. "The R&D community can absorb a cut of that magnitude for a year," he said. "The question is the duration of that kind of constraint."

Don I. Phillips, executive director of the National Academy of Science's Government-Industry-Research Roundtable, says, "Gramm-Rudman or not, the budget is going to be tight for several years." Mere belt tightening will not be enough, he warns. "What is necessary is a rethinking of R&D activities with the aim of finding how to do more with less."

## *In Challenger's Wake*

The explosion of the space shuttle *Challenger* has put in question not only the future of the U.S. shuttle program but also the fate of a number of important space-science experiments. Indefinite delays confront the 13 shuttle flights that had been planned for the remainder of this year. When a launch-pad fire destroyed *Apollo 1* in 1967, killing the three astronauts aboard, 18 months passed before the manned space program resumed. If as much time elapses now, 20 or more missions could be postponed or canceled.

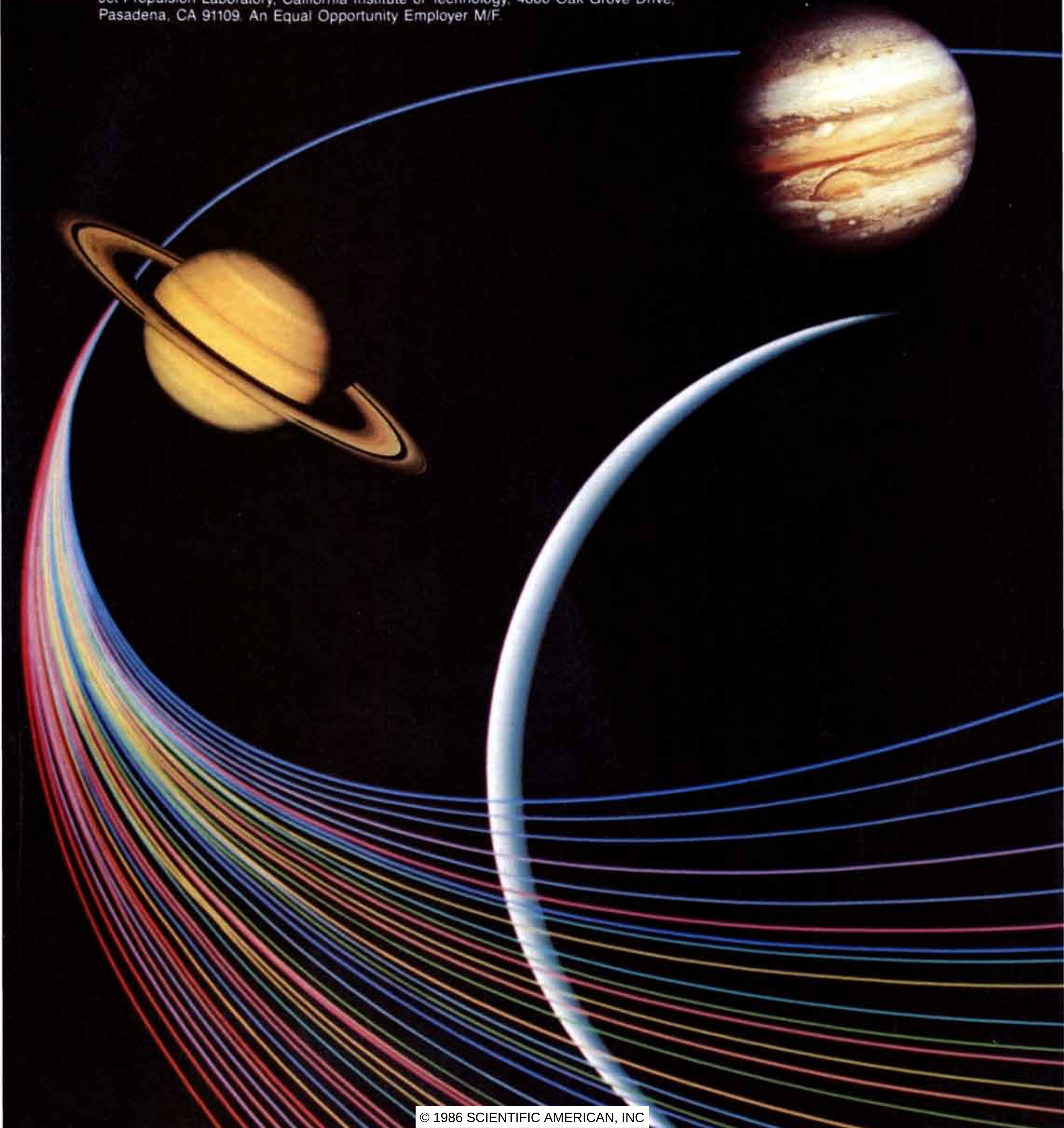
Several of the scientific missions face critical time constraints. Examples include the *Ulysses* and *Galileo* probes, which respectively are to explore the polar regions of the sun and the atmosphere of Jupiter. The spacecraft were to have been launched in mid-May; they will not be able to carry out their assignments again until June, 1987. Since the maximum separation of the two launchings is roughly five weeks, the remaining shuttle fleet (*Columbia, Discovery* and *Atlantis*) might not even be able to accommodate both missions in any given year. Either *Galileo* or *Ulysses* might have to wait until 1988 or later.

The Hubble Space Telescope was scheduled for a launch in October and may also be affected. The telescope would detect electromagnetic radiation ranging from the far ultraviolet through the visible to the far infrared. Since the instrument would orbit the earth, the timing of its launching is not so critical.

Because of their size and weight, *Ulysses, Galileo* and the telescope were designed to be launched from the shuttles. Little can be done, therefore, to expedite their missions. It is possible, however, that many of the commercial satellites scheduled for launching by the shuttle fleet could be placed in orbit by other means. Unfortunately the likeliest candidate, the European *Ariane* launcher, is booked for the next two years. Even conventional unmanned rockets require roughly three years to build. Over the next few years private U.S. launching companies might be able to take up some of the slack, but it seems unlikely that they can move soon enough to fill the void left by *Challenger*.

The military will also suffer. It is believed payloads carrying equipment related to the Strategic Defense Initiative were slated for July and August, and a photoreconnaissance satellite was to have been put in orbit in December. Although two years ago the Department of Defense began developing expendable rockets capable of carrying the large payload of a shuttle, none of these will be ready before 1988. If and when the shuttle program resumes, the Pentagon can (by a policy set by President Reagan in 1982) "bump" civilian payloads if it chooses.

Close to a year ago the National Commission on Space was directed to prepare a national agenda for space



exploration. Its report is to be presented on April 11. In the light of the *Challenger* disaster, few would be surprised if the commission recommends what some experts already argue: the shuttles should be employed only for the missions they alone can perform, and conventional rockets should be used to launch everyday satellites.

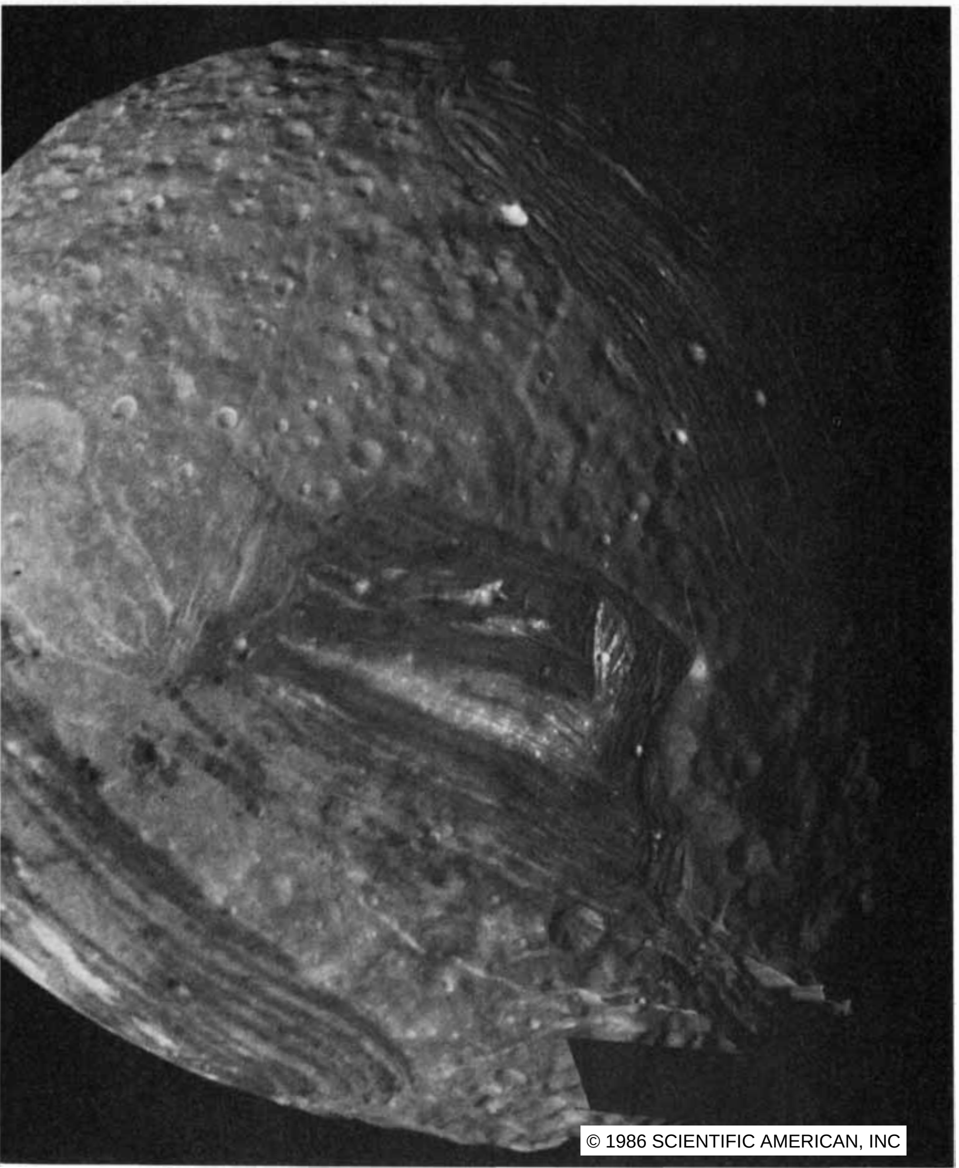
## *Encounter*

Anyone grown wary or even contemptuous of human technical achievements may yet feel a sense of wonder at the events of January 24. On that day a spacecraft launched from the earth nearly a decade ago flew by Uranus—Uranus, two billion miles away, invisible to the unaided eye, discovered by telescope only in 1781—and sent back a raft of surprising information that will occupy planetary scientists for years to come.

Prominent on *Voyager 2*'s list of surprises was Uranus' magnetic field. The axis of the field is tilted by 55 degrees with respect to the planet's rotational axis, whereas on other planets the tilt is less than 20 degrees. The rotational axis of Uranus is itself oddly tilted: the planet lies on its side, with the axis almost in the orbital plane. (At present the south pole points toward the sun.) So far no clear link has been established between the two anomalies.

On the other hand, the magnetic field has indirectly provided some clues to the planet's internal structure. One of the most widely accepted models predicted Uranus would have three layers: a molten rocky core about 8,000 kilometers in radius; a 10,000-kilometer-deep ocean of water, methane and ammonia, which are comparatively heavy volatile substances, and a 7,000-kilometer-deep atmosphere of molecular hydrogen and helium, which are light. According to Andrew P. Ingersoll of the California Institute of Technology, it now appears that the distribution of mass may be more homogeneous than had been thought.

The evidence comes from the planet's rotation rate, which has been calculated from Voyager measurements of the wobble of the magnetic axis about the rotational axis. The rotation rate together with the distribution of



**MIRANDA, innermost of the major Uranian moons, was photographed by *Voyager 2* from 31,000 kilometers away at a resolution of 600 meters (*top*). The troughs are several kilometers deep; white material is probably water ice. A computer-assembled mosaic of the moon (*bottom*) shows curious groove patterns (a chevron near the center and oval "racetracks" at the limbs) and a 20-kilometer-high cliff at the bottom right.**

## Buick redefines the Great American Coupe.

Once upon a time, America's roads were graced by handsome, big, sporty coupes. But in recent years, most big coupes grew to look suspiciously like sedans with two of their doors missing.

Introducing the new, front-wheel-drive Buick LeSabre Coupe. This is a whole new kind of big, sporty coupe. It rides and handles with a tightness and a crispness unexpected in a car of this size.

Its bigness will impress you immediately, for the LeSabre Coupe offers a rare commodity in sporty coupes: room. Genuine room for six.

Being a Buick, the LeSabre Coupe offers all this room in true comfort, and in plush surroundings.

As for its sporty side, it is an automobile truly designed for those who love the open road. The standard powerplant is a multi-port, fuel-injected 3.0-litre (not available in California). Or, order the available 3.8-litre with sequential-port fuel injection and roller lifters.

To handle this performance with great agility, the LeSabre Coupe has MacPherson strut front suspension, power-assisted rack-and-pinion steering and fully independent rear suspension.

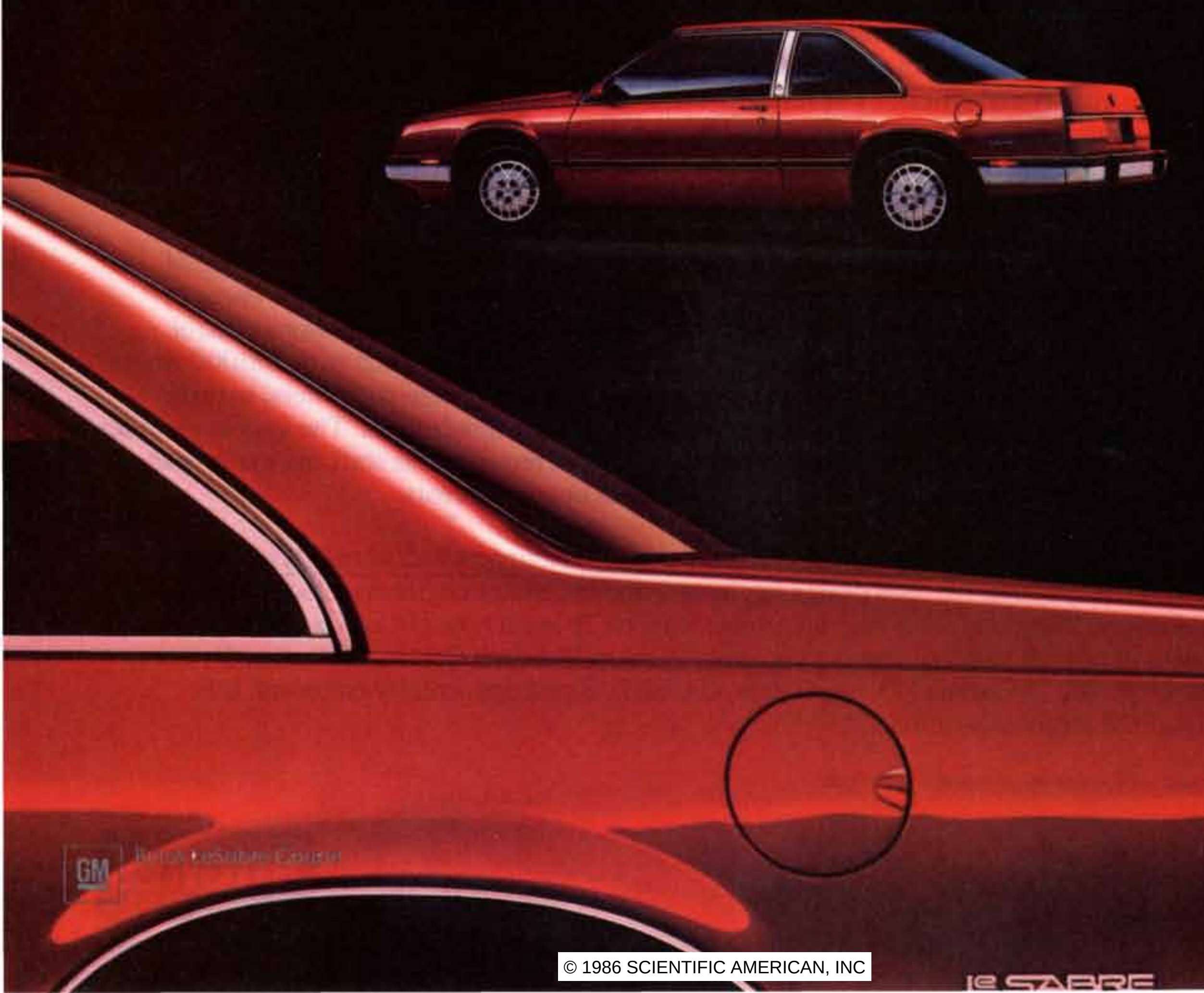
But enough of logic. Buckle up and visit your Buick dealer.

And experience the return of the Great American Coupe — in a very enlightened form. For more information, a test drive or a brochure, call:

1-800-86-BUICK (1-800-862-8425).

Wouldn't you really rather have a Buick?

BUICK



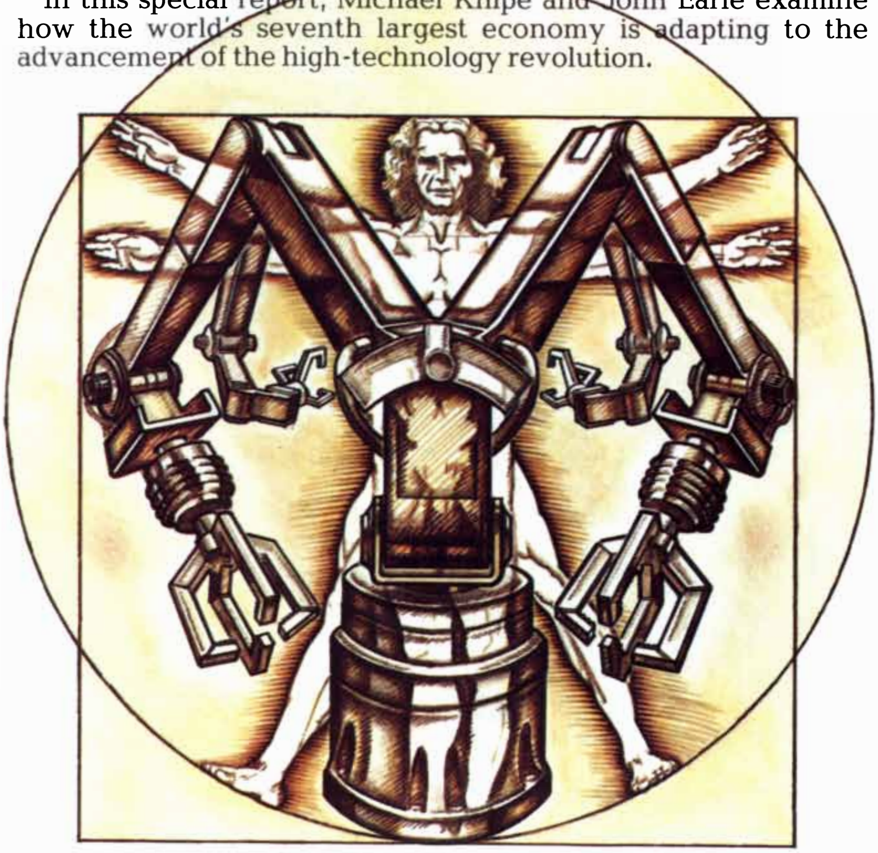
# ITALY

## INDUSTRY AND TECHNOLOGY TODAY

A steady resurgence in Italy's industrial self-confidence, which began at the beginning of this decade, has been particularly marked during the past 18 months.

Scores of companies have been tackling the problems of overmanning and have been investing in the replacement of outdated machinery with new high-technology equipment.

In this special report, Michael Knipe and John Earle examine how the world's seventh largest economy is adapting to the advancement of the high-technology revolution.



Italy's achievements are well illustrated by the success of the country's flagship industrial giants.

Olivetti has moved from losses and low productivity to become the biggest European-owned data processing manufacturer and one of the continent's most profitable large companies.

Fiat Autos, after losses of L.130 billion in 1980, announced profits of L.235 billion for 1984 and the Fiat group has embarked on a three-year program of research and development (R&D) that will total L.9,100 billion by the end of 1987.

Montedison, Italy's largest chemicals group, is expected to record a net profit in 1985 after reducing losses of L.393 billion in 1983 to only L.36 billion in 1984.

Italtel, the country's largest telecommunications manufacturer, has changed a 1981 L.268.8 billion loss into a L.25 billion profit and launched a prestigious new digital exchange system.

These financial turnarounds are particularly impressive following, as they do, the years of high wage inflation, industrial unrest, eroded foreign markets, massive losses and political terrorism that marked the 1970's.

But industrial growth in Italy is not only characterized by the major companies. Significant also are the hundreds of small and medium-sized enterprises. After the oil crisis of the 1970's – which hit the industrial structure of Italy particularly hard because of its heavy reliance on imported fuel – small and medium-sized manufacturers experienced a boom.

A striking element of this was the growth of new and original businesses which were independent of both the major capitalist companies and the state-controlled structures.

"We have witnessed," says Rome economist, Professor Andrea Saba, "an assimilation of capitalistic ways of producing with a uniquely Italian renaissance of craftsmanship, creativity and taste. You cannot understand Italy without taking into account Valentino, Pucci, Gucci and Benetton, and these are uniquely Italian companies springing not from the main industrial centers of Milan or Turin but from Florence, Bologna and Venice." Along with these well-established and world famous companies, the past decade has seen the development of others, less well known but equally adept at capturing today's markets.

The effectiveness of such companies results from the fact that they can respond quickly to market demands. They do not need to maintain a very complex distribution network. If the dollar is up, they can target their goods on the U.S.; if it is down, they move them to West Germany.

Says Professor Saba: "Consumers, when they are rich, are not influenced by cost but by taste and by variation of taste. Italy's strength is its ability to produce beautiful things. And consumers around the world, rich consumers, want to have beautiful things. If we produced shoes in competition with the Third World, we would be destroyed because wages in the Third World are one-sixth of wages in Italy. But if we produce beautiful shoes, every rich consumer will want to have Italian shoes, because they are Italian and much better than anything to be found elsewhere."

Italy's strength in this respect is surely the result of an innate sense of artistic originality born of 4,000 years of producing beautiful things.

"There are a lot of small enterprises in Italy where levels of competence are extremely high," says Professor Saba. "We have a good tradition of

brilliance in mechanics. From this has grown a kind of marriage between American electronics and Italian mechanics which is generating a lot of innovative numerical control tools.

"New technology does not only involve increasing efficiency. It also involves a reduction in dimension. In the past, merely because of size, only big concerns could possess a computer. Now there are mini-computers which are ideal for small businesses. So, in central Sardinia – a land of shepherds – you can find a small company producing traditional cheese which uses computers for controlling quality and which exports half its product to the U.S. In Sicily, you can find a small company using a computer in the production of furniture in a way I have never seen in any other part of the world. And there are other examples, including companies which are exploiting developments in biotechnology in the production of spaghetti, olive oil and wine." Prof. Saba is president of the Instituto per L'Assistenza Allo Sviluppo del Mezzogiorno, (the Institute for Assistance in the Development of Southern Italy).

In this respect, he stresses that while, in development terms, the south is perhaps five years behind the rest of Italy, it has considerable potential for attracting foreign investment capital. It is strategically situated in the center of the Mediterranean area, close to the markets of North Africa, as well as southern Europe, it offers relatively skilled and inexpensive labor and has a stable environment for industrial activity.

Throughout industrial Italy during the past few years, low inflation and a relatively strong export market have improved profit margins, providing funds for much needed capital investment. A great deal of this has been directed towards the automation of production processes and control systems.

A good example of this can be found at Benetton, the fashion clothing manufacturer with a worldwide network of owned and franchised retail outlets.

The biggest Benetton warehouse in Europe, with a capacity for 200,000 boxes of garments, is at Treviso, in northern Italy, where Comau, the Fiat group's innovative production systems company, has installed a computer-controlled handling and storage system that can store or remove boxes at a rate of one every four seconds.

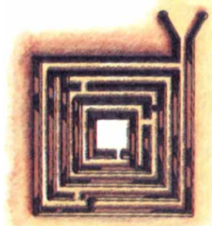
Similar advanced data processing and factory automation techniques are being introduced by a variety of other small and medium-sized Italian companies.

Although Italy's emergence from recession lagged behind that of most other countries, it has registered faster growth than any other country in Europe, with the exception of Denmark. Following an increase in Gross Domestic Product of 2.6 percent in real terms in 1984 – the best performance for a decade – a growth of 2.4 percent was expected last year.

The new expansion has not been accompanied by an acceleration in domestic inflation. In 1984, this averaged only 10.6 percent compared to 15 percent in 1983, substantially reducing the wide gap in performance which had previously separated Italy from other industrial countries.

If Italy is to maintain its economic recovery over the long term, however, banking sources stress that the government will have to be much more vigorous in its pursuit of lower public expenditure and in its methods of keeping wage increases down.

## R&D POLICY



One of the strengths of Italian industry has always been its flair for both entrepreneurial and technological innovation. However, this has all too often lacked the backing of a broadly conceived national strategy for the development and application of new technologies.

The 1980's, however, have seen a marked increase in the degree of priority placed by the government on scientific and industrial-oriented research.

Italy spent altogether L. 8,575 billion on research last year, according to the official statistics bureau. This is well below the two percent of Gross Domestic Product regarded as minimum in most industrial economies, but it represents a 12.2 percent rise in monetary terms on 1984. When purged of inflation, the real increase is about four percent.

The government and public authorities were responsible for L.3,983 billion, nearly half the total, while the corporate sector spent the rest. Public expenditure was divided under three headings: L. 1,442.6 billion on pure research; L. 1,402.5 billion on applied research; and L. 137.8 billion on development research.

The Ministry of Scientific and Technological Research suffers from the drawback of being headed by a minister who is officially without portfolio.

He is dependent on the office of the Prime Minister who delegates to him responsibility for scientific research.

The Ministry is given powers of supervision over bodies such as the National Research Council (Consiglio Nazionale delle Ricerche or CNR), and much of its work depends on maintaining what Rocca calls relations of collaboration and trust with other ministries, particularly defense, industry, agriculture, health and posts. Despite the absence of cut-and-dried regulations, the system works.

In its task of overseeing and stimulating industrial-oriented research, says Rocca, the Ministry follows four paths.

The first and oldest is a procedure going back to a law of 1968 under which the government provides soft loans and grants for research projects initiated by companies. Originally a L. 100 billion fund was set up under the administration of Istituto Mobiliare Italiano (IMI), a finance house owned partly by banks and insurance companies. The legislation has since been modified and added to several times.

The most recent law is dated February 1982. It provides for the Ministry to preselect applications from firms which are then passed to IMI for detailed examination while a committee of ministers determines the amount of public financing to be advanced. In its annual report for 1984, IMI said it had been asked by the Ministry to examine 166 projects for an overall estimated cost of L. 513.4 billion. At the same time, 116 new research contracts were signed for a value of L. 420.1 billion, of which L. 314.3 billion were advanced from public funds.

The 1982 law was a milestone because it opened the gate to the second path – government-initiated research.

"Instead of industry asking for public financing," Sig. Rocca said, "this is now a case of the state asking industry to undertake research in areas of national importance, which would not give firms an immediate return."

The third path is space. A national space plan exists and has been put under the temporary supervision of the National Research Council. Space research should take off as a result of a cabinet decision in August to set up an Italian Space Agency (Agenzia Space Italiana).

An important bilateral program, cited by Sig. Rocca, is underway with NASA in the field of satellite telecommunications. In the multilateral field, Italy participates in the European Space Agency's Columbus program, while it is keenly interested in both the U.S. Strategic Defense Initiative and in the French-inspired Eureka project.

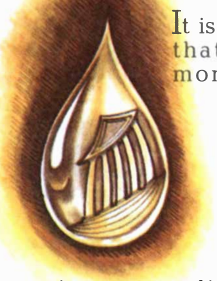
The fourth path leads to so-called "application-oriented projects" (progetti finalizzati) of CNR, which receive government funding and are supervised by the Ministry. Their advantage lies in the way, new to Italy, in which they bring together public bodies, universities and industry. The program began only 10 years ago and got properly underway after 1980. And as the average length of a project is five years, the results so far are limited. By this year, said Sig. Rocca, 28 projects had been completed while 11 were underway and two were starting.

Much of the progress is due to the CNR's decision to set up an Office for Technology Transfer in July 1980, headed by Dr. Sergio Allulli, a chemist. Before then there was an enormous gap to be bridged, says Dr. Allulli. "Italian universities knew little about the value of a patent, nor did people want to make the results of research comprehensible to firms. The researchers themselves were not convinced that what they were doing was really of use to industry."

On the other side, he said, much of Italian industry "feared innovation because it is a risky business and the state did not create the proper climate. The risk was very high and we do not have venture capital here."

"We don't need to spend more on research," says Dr. Allulli, "we need to utilize it better." In previous years "we wasted time – the results remained shut up in drawers. It was as if Fiat made its cars and then shut them in warehouses."

## FINE CHEMICALS



It is perhaps appropriate that Italy's classical monuments, having suffered the polluting effects of 20th century technology, should now be able to enjoy respite thanks to a sophisticated fluorinated fluid whose primary application is in the aerospace, electronics and nuclear industries.

It is also fitting that this fluid should have been developed and marketed by an Italian company. Thus, the country's rich architectural past is linked with its leading role in state-of-the-art technological development.

The monument-saving fluid is Fomblin Y Met, a perfluoro-polyether which is produced by Montefluos, a leading producer of fluorine derivatives and a part of the giant Montedison group which during its 100-year history has introduced numerous innovative products including ammonia and polypropylene.

The deterioration of Italy's architectural treasures is caused mainly by pollutants in the air, which react with the carbonate part of the stone in the presence of water.

The remedy is to eliminate or reduce the penetration of water in the stone, while ensuring that the protective measures do not introduce any other destructive activity. Fomblin Y perfluorinated fluid is unique in that it is the only water and oil repellent material produced on a commercial scale that is chemically inert, resists the most aggressive agents such as chlorine, fluorine, acids and alkalis; is insensitive to extremes in temperature; and is stable and transparent but permeable to air and water vapor, thus permitting controlled aeration and humidification of the internal layers of stone.

Among the monuments that have now been treated are the Palazzo Pitti in Florence, the Apollo statue in the Boboli Garden in Florence, the Duomo di San Martino in Lucca, the Palano Ducale of Urbino and Cappella Paszi of St. Croce in Florence.

"The effect of Fomblin is to provide protection without giving a coating effect," says Dr. Paolo Bolzani, managing director of Montefluos. "Fomblin is a technology-intensive product and it is expensive. But its properties are truly exceptional and they make it indispensable for use over a vast range of technologies such as the nuclear, electromechanical, aerospace, magnetic media and semiconductor industries.

Dr. Bolzani believes that Montefluos can use its strong position in perfluoro polyether technologies as a building block in the search for new advanced materials. "We have an exchange of ideas with our customers in which to say, 'forget about current technology; if you can envisage a completely new compound with a combination of performances which is completely new, what would be required of us?' Starting from this point we expect to develop a new generation of materials which cannot yet be named." Elaborating on this theme, Dr. Gerardo Caporiccio, the company's R&D manager, says: "Our products now work across a wide temperature span, from – 90° centigrade to + 250°. But what if we can consider a span between – 120° and + 320°? There is no reason why we shouldn't eventually achieve this. Our strategy is to move, conceptually, the frontiers of expectation of our customers."

Montefluos' sales increased by 25 percent last year to reach $233.8 million, with sales of Fomblin fluids increasing by over 50 percent.

Montefluos' success is indicative of the remarkable turnaround that has been achieved by its parent company, Montedison, which, in addition to its chemical interests, has health-care and energy subsidiaries.

## DATA PROCESSING

The development of a voice-activated workstation is a prime, though still distant, goal of the office automation and data processing industry.

Its realization will require an extremely complex and sophisticated system which will have to include linguistic and grammatical knowledge and interpretative capabilities.

This is an area of particular interest to Olivetti because of its role as Europe's leading office automation and data processing manufacturer. With more than 70 percent of its sales on the continent, it needs to solve the problem simultaneously in at least five languages – Italian, English, French, German and Spanish.

"Precisely because of this we are intensely involved in the area of speech recognition," says Dr. Arnaldo Pasini, vice-president responsible for corporate research.

The state of the art in computerized speech recognition is the achievement of simple comprehension of vocabularies of between 4,000 and 6,000 words, still far short of the 15,000 words used daily by most people. Beyond the problems of acoustic recognition and vocabulary building, however, lie the difficulties of machines understanding uninterrupted speech in which words and phrases run into one another, and the complexities of semantic analysis.

Another and shorter-term aspect of the office of the future which is under study at Olivetti is non-impact printing. Easily accessible color imaging is expected to gain increasing importance as a practical means of improving the information content of office documents. Among the technologies being explored at Olivetti are thermal ink transfer, electrostatic, magnetic and ink-jet printing.

Thermal transfer and ink-jet have the advantage of permitting the use of plain paper. In digital printing, color is obtained by means of patterns of elementary dots of variable filling, with minimal intervals of one dot, according to the level of saturation required, so that the number of pixels of color per unit of length is smaller than the corresponding value for monochrome printing.

However, this does not imply a lowering of quality because color is added information that well outweighs the decreased resolution.

Thermal transfer can be carried out at reasonably high speed, with great flexibility and a variety of colors. The technology will enable a color printer to reproduce, in a few seconds, color copies of good quality with a range of shades and several different colors.

One difference between thermal transfer and ink-jet reproduction is that thermal transfer provides slightly brighter colors while ink-jet offers a wider range of intermediate shades.

A bigger difference, however, is presented by the economic factor. Thermal-transfer technology will be relatively simple so the cost of the machine would be small but the cost per copy of reproductions would be high. As ink-jet technology is more complex, the initial cost would be higher but the cost per copy would work out much cheaper.

Dr. Pasini believes it will only be a few years before speedy and economical full-color reproduction will be available as a regular item of office equipment in much the way photocopiers are today.

"Personal computers will have a color screen and people will want the facility of making hard copies in color. What is now under investigation is how to reduce the complexity of the copiers while retaining quality and speed."

All of the new technologies being researched are subjected to the significant constraints of costs. Developments may be technically superb, says Dr. Pasini, but the final, critical factor is whether they can be successfully marketed at a price acceptable to the average business office.

Olivetti has continually enhanced its products against stiff international competition by increasing its investment in research and this has been a significant factor in confirming its leadership in Europe. And having attained this position, the company is maintaining its growth path by a double strategy, according to Carlo De Benedetti, the company's charismatic chairman and chief executive.

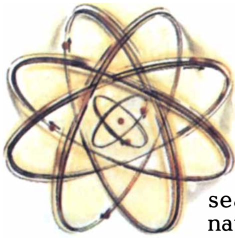
Olivetti internal development efforts may be summarized as follows. In the last five years the Group has invested approximately L.770 billion in R&D. Thirty-two hundred people are engaged in R&D mainly concentrated in the Ivrea (Italy) area, at the company's headquarters. R&D personnel account for 6.8 percent of the Group employees and for 10.4 percent of employees in Italy. R&D expenses in 1984 rose by 22 percent compared to 1983. Besides the Italian laboratories, Olivetti has R&D centers in the U.S. (Cupertino, California), France, Switzerland, Spain and Singapore.

An example of the results of the R&D design and production efforts at Olivetti is the M24 Personal Computer, whose favorable acceptance even in the most demanding markets like the U.S. is demonstrated by the $100-million contract with AT&T for supply of the M24 to be marketed by AT&T under their own label, and by the contract signed with Xerox for OEM supply of the M24.

The communications and data processing sector continues to be marked by a global restructuring of international suppliers. The factors behind this are the rapidly emerging technological developments within the industry, the convergence of telecommunications with data processing and the deregulation of telecommunications services.

The availability of the basic telecommunications technologies developed by Bell Laboratories, one of the largest research centers in the world and easily accessible to Olivetti, is regarded as one of the essential elements for the "systems integrator" function demanded by modern office automation.

## ENERGY



Italy must meet more of its energy needs by importing oil than any other major industrial power and the search for alternative sources has therefore been a major priority.

In the national energy mix, oil now accounts for less than 60percent of total consumption, against 67.2 percent in 1980, and is targeted to fall to 53.7 percent by 1990. Natural gas consumption, now 19.3 percent, will increase to 20 percent, and solid fuels, 10.3 percent now, should rise to 13.5 percent.

ENEL, the National Electricity Utility, optimistically said in 1975 that 50 percent of electrical power would be of nuclear origin by 1985, and more than 80percent in 1990. In fact, the percentage is only 3.2 now and under the revised national energy plan, will be 7.2 percent in 1990. The nuclear program lags far behind its original schedule, partly because of the fall in world oil prices and partly because of strong local opposition to new plants.

Currently, nuclear power comes from the 850MW Caorso plant (BWR, under General Electric license) on the Po river and two obsolescent units, Trinto Vercellese (270MW) in Piedmont and Latina (200MW) south of Rome. Another BWR plant of 2,000MW (twin units of 1,000MW) is under construction at Montalto di Castro, north of Rome, and three PWR plants of 2,000MW to be located in Piedmont, Lombardy, and Puglia are at various stages of planning.

The future beyond that is the province of ENEA, the Commission for Nuclear and Alternative Energy Sources. ENEA has a project to develop a standard Italian PWR reactor of 2,000MW, and a 40MW prototype thermal neutron reactor is undergoing preliminary testing near Latina, which is intended as a first step towards the realization of an alternative type of plant.

A budget of L. 1,760 million is being provided for work on fast breeder reactors. This includes a third share in the French Superphènix program, as well as the Provoa Elementi di Combstitili (PEC) project underway at Barasimone in the Apennines south of Bologna. This is a 120MW fast breeder reactor with a central irradiated channel for testing fuel elements in conditions of industrial application and is due to be charged with fuel in 1988.

ENEA's overall policy is aimed at making Italy independent of foreign licenses and competitive in international markets.

*The nuclear power station at Latina, south of Rome.*

Ferrara, an historic city of 150,000 inhabitants near the mouth of the River Po, is to be centrally heated by geothermal energy as part of the country's campaign to lessen its dependence for energy on imported oil.

A distribution network which is being installed will reach 5,000 civil and industrial users in its first phase, but the program calls for the eventual substitution of Po Valley methane by the cheaper geothermal system throughout the city.

The wells are about two miles from the city center. The fluid, water at nearly 212°F, with a high saline content, lies at depths of between 4,000 and 6,100 feet. Because of the corrosive effect, the thermal energy has to be transferred to a second, freshwater circuit for distribution, while the fluid in the primary, closed circuit is reinjected into the subsoil. To satisfy demand at peak periods, conventional boiler heating will reinforce the geothermal flow of 400 cubic meters (about 14,120 cubic feet) an hour.

Agip, the chief operational subsidiary of the state hydrocarbons corporation, ENI, is in joint venture in this and similar initiatives with ENEL. Agip entered the field because geological, seismic and drilling know-how is very similar to that of oil exploration.

Another agreement was signed in February 1985 with Vicenza, a city with a population of 110,000 in the Venetia region and birthplace of Palladian achitecture, where fresh water at 152°F has been found at a depth of 6,000 feet inside the town limits. The agreement provides for delivery to begin in March 1987, at the rate of 120 cubic meters (about 4,238 cubic feet) an hour, for heating public and private buildings with a volume of about one million cubic meters (about 35.3 million cubic feet).

Plans are also underway for the ENI group's headquarters at San Donato, near Milan, to be heated geothermally. In the commercial field, an ENI subsidiary has 8,121,000 square feet of geothermally heated greenhouses for flower cultivation on the slopes of Monte Amiata, an extinct volcano in southern Tuscany. It is believed to be the biggest orchid garden in Europe.

At present, there is a gap in legislation on how geothermal resources can be developed. Agip is pressing for the speedy passage through parliament of a bill which would fill the gap by opening exploration to private as well as public operators, both Italian and foreign.

## AUTOMOBILES

An innovative car with a fuel consumption of 112 miles per gallon has been developed by scientists and engineers at the Fiat Research Center in Turin.

With five passengers, it can cruise at 90 miles an hour and accelerate from zero to 437 yards in 20 seconds. Its shape has been dictated by aerodynamic principles and wind-tunnel experiments. Its back suspension is provided by laser-controlled pneumatic springs. It has two extremely wide doors, each on a double hinge, to provide easy access to the rear seats without the need to move the front ones. It is constructed of steel and reinforced plastic and weighs only 220 pounds.

It is, of course, a hand-made experimental prototype, whose purpose is to provide components that may eventually become innovative commercial products in one form or another. One imminent prospect, in this respect, is a 1.7 liter direct injection diesel engine. The problems of noise and smoke are being resolved and a 15 percent lower fuel consumption is being maintained.

At the Fiat Research Center, says Dr. Angelo Rossi, a special assistant to Fiat's chairman, Gianni Agnelli, the emphasis continues to be on cutting fuel consumption and improving vehicle quality, reliability, safety and comfort. Considerable emphasis is also placed on the development of automated production process and control systems as a means of enhancing product competitiveness and quality.

The Center has achieved some outstanding results in manufacturing technologies. These include an engine cylinder surface monitoring system, the prototype of a welding spot-checker, a laser system for a two-station welding plant, a complete computerized system for automated component test benches and a control booth for dynamic diesel engine torque tests.

It has also intensified its work on the development of new sighted systems for use in robotized plants. The sighted robots would be capable of selecting mechanical components and would also be employed in laser applications for the antiwear treatment of dies and sheet cutting.

Fiat is at the forefront of laser beam research. Lasers are being used to study the behavior of mechanical structures under vibration or stress and also in production to check the surface roughness of crankshafts. At Fiat Auto's Mirafiori factory in Turin, lasers are used to weld gearbox synchronizer rings and to inspect cylinder surfaces.

The epitome of Fiat's advanced technology is the Fully Integrated Robotized Engine (Fire), which was unveiled in September 1984 and last year fitted into several basic Fiat models.

The Fire 1000 engine is the forerunner of a new generation of Fiat Auto power plants for one-liter cars. It is based on highly advanced R&D techniques and reconciles the objectives of low weight, high efficiency, low specific fuel consumption and totally automatic high-output production.

The most obvious difference between the Fire 1000 and its predecessors lies in the reduced number of components – 273 as opposed to 368 for the 903 cc engine. The next big difference is the compact size and greatly reduced weight, a mere 69 kg. The engine block on its own weighs 18 kg, six kg less than the 903 cc engine.

A new casting system is being used in the manufacture of intake manifolds. Called Policast, it uses disposable expanded polystyrene models. Innovative design and manufacturing techniques have also permitted the application of cast iron. This material has been used with such great success that the engine block, caps, drive shafts, conrods, camshaft, flywheel valve guides and piston rings are all in cast iron. Engine block walls are no thicker than 4 mm. The crankshaft weighs 5.93 kg, whereas the steel 903 cc engine's crankshaft weighs 7.4 kg. The light alloy cylinder head is made in one piece.

Construction and assembly operations have been simplified to such an extent that the Fire only takes half as long to make as the 903 cc engine.

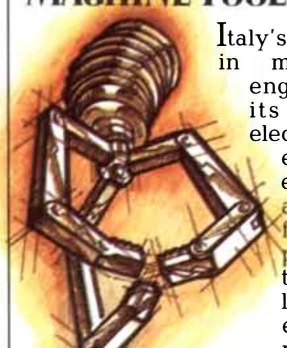
For the first time, the production workshop has been designed hand-in-glove with the new engine it is to produce. As a result, there are five parallel production lines, an intermediate store to ensure constant supply of materials and a flexible production rate of between 1800 and 2600 engines a day. The plant controls itself with computers, runs checks on itself and submits its products for approval by itself, all automatically. Every operation is computer controlled. Engine flywheel turning takes place, for example, on an automatic transfer machine which is considered to be one of the foremost unmanned workshops in Europe and still the only transfer machine of its kind in existence. Not only are operations carried out with numerically controlled lathes, but tools are automatically replaced under the direction of sensors which keep a constant check on machined components. The transfer machine also includes a robot which loads the unmachined flywheels using a sonar-fitted magnetic clamp. This device helps the robot search for flywheels in a bin of loose components. Production machines are fitted with diagnostic and monitoring systems to optimize line efficiency and productivity and to keep a constant check on product quality.

It takes just 107.5 minutes to make each engine and one comes off the line every 20 seconds. Cutting parameters are adjusted and idle times are reduced to ensure that all machines are utilized to their maximum capacity. The head attachment side of the engine block is milled in 0.34 minutes, whereas the same operation for the Ritmo 138 engine takes 0.65 minutes of machine time. Similarly, rough turning of crankshaft pins is now carried out in 0.382 instead of 1.20 minutes. The burnishing of the small conrod end takes 0.19 instead of 0.62 minutes. Machining takes only 46.5 minutes instead of 114.4 and assembly is completed in 61 minutes instead of 117.1.

These seconds shaved off every operation mean that the Fire 1000 is made in less than half the time it takes to make the Fiat 127 903 cc engine.

Robotization and the development of on-line process computerization and automated control systems – as exemplified in the production of the Fire 1000 – are the keynote elements of Fiat's automobile sector production. The company spent L. 266 billion on R&D last year, a 13 percent increase on the previous year, and this amounted to 2.1 percent of revenue. Fiat Auto's net indebtedness was reduced by L. 547.6 billion and the work force was reduced by 13,566 to a total of 107,681. Production was up 2.8 percent and totalled 1,409,700 cars.

## MACHINE TOOLS



Italy's preeminence in machine-tool engineering and its advanced electronic control expertise has established it as the world's fifth largest producer and the fourth largest exporter of flexible manufacturing sytems.

Machine-tool exports total nearly L. 1,000 billion annually. Continued growth, however, depends on the ability of manufacturers to further integrate computerized electronic control with machine-tool technology and the development of more advanced flexible manufacturing systems.

Comau, the Turin-based production systems company in the Fiat group, has achieved particular success with its innovative robotics development and with its increasing integration of mechanical and electronic technology in its metal working and assembly systems. Sales increased during the 1984 financial year by 21.4 percent to L. 498.2 billion.

The company is continuing to extend the range and versatility of its Smart robot, an electromechanical articulated machine with six degrees of freedom, which can be adapted to perform a variety of tasks.

A process for the automatic assembly of hinges on motor vehicle doors has been developed by equiping a Smart robot with a telecamera and a microprocessor. The robot takes a pair of hinges, complete with bolts, from one pallet. Once the bolts are in line with the holes, the robot screws them down. The tightening torque and final rotation angle are checked bolt by bolt to ensure that each hinge is properly assembled. A coded label is applied automatically to the door if the operation is incorrectly performed. Up to three different types of doors can be handled at the same time.

Another element in Comau's success has been its ability – at a time of uncertain markets and fluctuating demand – to provide manufacturers with innovative production systems capable of great flexibility, thus providing them with the facilities to fine tune their mix of products and to reduce the lead time from design to market.

Often the basic concern of manufacturers is to arrange their production scheduling so they can meet their order dates. This demands controlled order release, adequate capacity, control of the level of work in process, manipulation of job queues at each work center, minimizing machine and manpower idle time and preventing bottlenecks.

To meet these needs, Comau has developed a comprehensive software package to deal with operations control, planning, reporting, machine programing, in-process inspection, tools management, plant information system interface and plant materials handling.

As a result, foreign sales this year are expected to constitute some 60 percent of total revenues. Major orders have been placed by several factories in the Soviet Union, and in the U.S., Comau has sold one of its most complex flexible systems developed so far to Borg-Warner's compressor manufacturing plant in York, Pennsylvania.

The production is almost entirely automated. The only human intervention on the shop floor is the man who picks up the blanks delivered from the buffer stock and introduces them into the roller track.

## ENGINEERING

Coal transportation by means of slurry pipelines has been hindered by environmental concerns and a reluctance on the part of end users of coal to deal with the problems of dewatering it at the discharge of the pipeline.

Recently, however, a new concept of coal-water slurry with higher coal concentration has been pursued and Snamprogetti, an Italian engineering company specializing in energy infrastructures, has developed a new and more advanced version of this technology. It has devised a method by which coal can be converted into a liquid with a high coal concentrate (more than 70 percent), very good fluidity and stability. The liquid is suitable for conventional handling and burning as a liquid fuel without further processing. Potential users are large power stations and industrial plants, particularly those whose location is unsatisfactory for supplies of coal. Distribution can be by the same means as heavy fuel oil – by pipeline or large tankers.

A prototype plant with a capacity of about 100,000 tons is being built for Laviosa, an Italian company, and Snamprogetti has signed a contract with the Soviet Union for a plant to convert coal into slurry and a 120-mile pipeline to convey it to a power plant in Siberia.

Snamprogetti's sister company in the ENI state holding company, Saipem, has meanwhile developed what is described as the first fully automatic, fully computerized welding system suitable for marine operations.

Productivity in the oil, gas and water pipelaying industry, particularly offshore, is critically dependent on the speed and effectiveness of the pipe-welding operations. And after extensive prototype tests, including operational projects onboard Saipem's offshore pipelaying vessel, three of the computerized welding units are being manufactured and are expected to become operational later this year.

Snamprogetti's coal-water slurry and Saipem's welding system illustrate the degree to which dependency on imported oil has promoted Italian innovation in the design and manufacture of systems and equipment required to meet the country's energy needs. Italy must meet two-thirds of its energy requirements from imported oil, a higher proportion than any other major industrial country.

Snamprogetti and Saipem – both highly profitable companies that earn 85 and 77 percent of their revenues respectively from abroad – were partners in laying a 1,500-mile trans-Mediterranean pipeline to bring natural gas from Algeria across the Messina and Sicilian channels to Italy, a project which came on stream two years ago. It involved pipelaying at 2,000 feet, a greater depth than had ever been accomplished and was a major breakthrough in marine technology.

"Seabed technology is the technology of the future," says Dr. Duilio Greppi, Snamprogetti's chairman. "Today we are capable of laying pipes not at 2,000 feet but at 2,000 meters. We are doing a great deal of intensive research in seabed technology. Much of it must remain secret but the trend is towards having seabed level platforms that will be capable of separating oil from gas, burning off the unwanted gas and pumping only the oil through the pipes. We are also working on the development of technology to enable seabed operations in difficult geographic situations, such as the Arctic and Baring Seas. Another aspect of our research concerns improving the technologies for freeing natural gas from $CO_2$ and/or $H_2S$, which could open up many gasfields which at present are uneconomic."

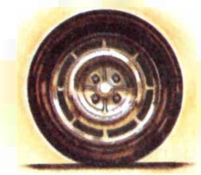
Snamprogetti is involved in consulting, engineering and prime contracting for oil, gas and petrochemical plants as well as for onshore and offshore pipelines.

Saipem has also established a widely recognized technical leadership internationally in deep water pipelaying. This has been due, in part, to its ownership of the most sophisticated offshore construction fleet. The *RAGNO DUE*, a comprehensively equipped 4,200-ton diving mothership is the first such vessel to carry two types of submersibles – a submarine and a tethered diving bell – offering clients the choice of both working independently or together. The *RAGNO DUE* provides a variety of functions: pipelaying, surveying, search and recovery, diver support and pipeline inspection.

Another vessel, a new-generation semi-submersible, is capable of working in up to 40-foot seas and laying pipes to a depth of 4,000 feet.

Says Sig. Dario Dall'Aglio, head of Saipem's external relations division: "Our experience has given us valuable insights in terms of operating limits and control accuracy which we believe are vital for safe, economical and effective operations at sea."

## CABLES & TIRES



With a wide ranging transformation of the telecommunications industry underway internationally, Pirelli, the Milan based power cable and tire manufacturer, is well set to maintain its leading role in the industry.

Having patented, in 1917, the first oil-filled cable – still the best means for power transmission at very high voltages – the company is now in the process of developing a new oil-filled cable which can be impregnated with special flame resistant synthetic liquids. It is also developing very high voltage submarine cables insulated by paper impregnated with a viscous mixture of fluid oil and capable of being laid at a depth of 6,500 feet. Other projects in progress include the establishment of an experimental plant for cables and accessories for use at 1,100kV, the highest voltage yet applied to a cable; research on optical submarine cables suitable for long distances and great depths; and the development of cables featuring a high degree of reliability when used in especially exacting conditions such as extreme heat, nuclear radiation or in situations where it is essential to reduce the emission of toxic fumes and gasses.

Cables account for 44 percent of Pirelli's turnover, with tires accounting for 42 percent and diversified products for the remaining 14.

In the early 1980s, Pirelli pioneered the development of tires with low rolling resistance which reduced the absorption of power thus saving fuel.

Says Dr. Franco Bottasso, Pirelli's director of tire R&D: "Before the energy crisis, a good passenger car tire had a rolling resistance coefficiency of around 13 kilos per ton. Today it is down to 10 kilos per ton. The question now is how much further it can be reduced. For a passenger car we could go down to six or seven but this would have to be done at the expense of performance, mileage, comfort and handling."

For the remainder of this decade at least, says Dr. Bottasso, tire development will continue to be strongly influenced by the need to reduce energy consumption and to compensate for the high price of raw materials.

"In general terms, innovation will be conditioned by the need to ensure that the product gives the necessary performance with the minimum weight and minimum power absorption. Different balances will have to be found from those accepted until now between load, deformation, fatigue resistance and running temperature. We can realistically forecast further fuel savings of between 3 and 5 percent. We will have to take a close look at the availability of the current raw materials and identify which other ones might offer more effective use in terms of cost and availability. This will involve new formulations and perhaps even the redesign of the tire. This, of course, presents a major challenge, given the fact that the tire is a remarkably effective engineering concept that has remained virtually unchanged for a century and a half."

More immediately, Dr. Bottasso foresees the introduction, on small economy cars used mainly as city runabouts, of smaller more highly inflated tires; and the introduction of a tire capable of continued mobility even when flat which would dispense with the need for a spare wheel. Another prospect, he says, will be the design, for trucks, of wide single tires to replace the twin tire combinations of today. "If you can produce one tire that can carry the load of two it could be lighter, require less material and dissipate less energy."

Today's tire research at Pirelli is directed towards further reductions in rolling resistance for even better fuel economy, the development of high performance tires for sports cars and competition, the development of truck tires with improved load carrying and speed characteristics and the investigation of new materials and tire constructions.

## TRADE FAIR

The dramatic advances in communications technology in recent years have had a profound impact on the role of international exhibitions and trade fairs. To adjust to the changing requirements and to utilize the latest equipment, a major three year reconstruction program has been launched by the Milan Fair organization.

"The role of great fairs in the future is clear," says Mario Boselli, the Fair's new president. "They must be transformed from simple exhibition complexes into active business promotion centers. By that I mean permanent and specialized trade centers capable of providing businessmen from all over the world with integrated consultancy and assistance services. Obviously, as the year 2000 approaches, a great international fair cannot just sell display space. It has to provide the latest communications facilities, marketing procedures and services."

In order to do so, the area occupied by the Milan Fair, close to the heart of the city, has been transformed into what Sig. Boselli calls an "optical island." At a cost of L.10 billion, a network of optical fibers has been installed to facilitate the transmission and reception of video telephone calls, provide video access to data banks and televisual link-up with various locations elsewhere in Italy and Europe for conference purposes.

If anyone should suppose that the increasing sophistication of communications technology means that there will be a declining need for international trade fairs, Sig. Boselli is quick to correct them.

"Fairs will always exist because of what you might call the moment of encounter. The need for specialists to confront other specialists, to examine each other's goods directly and to exchange information cannot be satisfactorily done by any other method.

"Visitors will still want to see goods on display, to examine the finished product. Nevertheless, the sophisticated use of video facilities will mean that exhibitors will be able to put on display their factory production processes."

Milan and its Trade Fair have become almost synonymous in the 65 years of the Fair's existence. The city itself has been a central European trading post for centuries, in spite of having no indigenous resources. Built on the plain of the Po River, 90 minutes from the port of Genoa and close to both the Swiss and French borders, its location has helped make it one of the continent's principal financial and industrial centers. Today it is the major trading base of most of Italy's industrial concerns, the home of such corporate names as Alfa Romeo, Montedison, Snamprogetti, Saipem and Pirelli. Two of Italy's three state-controlled banking institutions – Credito Italiano and Banca Commerciale Italiana – are based in Milan, as are 33 foreign banking institutions. It is also the focal point of the design and fashion businesses.

Together with the surrounding region of Lombardy, Milan accounts for a third of Italy's Gross Domestic Product and a third of the country's exports. It is therefore no wonder that the Fair continues to be a prime rendezvous for international trade and industry.

Sig. Boselli contends that it generates 25 percent of Italian exports and brings in L. 1,200 million worth of business annually.

The Milan Fair organization stages about 80 specialist exhibitions each year and hosts more than 300 other conventions. The biggest annual event is the Fiera Campionaria, the international exhibition which featured more than 7,000 stands last year, more than 1,800 of them from 79 foreign countries. Among them were Albania, Syria, Iran, Mozambique and Vietnam.

Under Sig. Boselli's direction, the Fair is in the midst of an evolutionary change aimed at stressing its function as a permanent center of active business promotion. "We expect the number of exhibitors each year to decrease but the degree of specialization to increase," he says. "Our first task is to increase the available space for producing exhibitions and improving the facilities for mounting and dismantling large and complex exhibits speedily. At present, the exhibition area is 600,000 square meters and there are 43 halls and pavilions. An additional 50,000 square meters of open space is now being developed."

The new telecommunications and data processing facilities have been designed and installed by Italtel in collaboration with SIP. A network of optical fiber cables totalling nearly 400 miles in length connects a production center with the exhibition stands. Over this network the simultaneous transmission is effected of voice, text, data and images, controlled by an Office BX 100 exchange, which is a satellite of the private digital Office BX 5000 of Italtel Telematica.

The Fair's Service Center is comprised of a production room which controls the distribution of all video programs and video telephone connections between users and an area housing all the systems. The production room has the local source of pictures, tuners for the conversion of television signals via radio to video signals, a set of monitors and an operator station. The equipment room houses an Office BX 100 central unit, modems for transmission of data, a PDP 11 mini-processor which checks and organizes the connections between subscribers and the video telephone service and the selective distribution of the available video programs; a wide-band switching matrix which permits the transmission of video signals between subscribers and between them and the service counter; and a transceiver which converts electrical signals into optical signals.

The terminals used for the interactive television service also perform the role of a video telephone enabling color video pictures to be received and transmitted simultaneously with telephone calls.

Says Sig. Boselli: "Exhibitors and visitors at the Fair will have at their disposal both advanced telephone services and new telematic services. This optical island project is an experiment, an opportunity for research and verification which should break new ground in developing exhibition facilities. Visitors to the Fair will be invited not only to see for themselves what communication technology offers and how it operates, but to participate themselves in the experiments."

**ITALY: INDUSTRY AND TECHNOLOGY TODAY**
**Editor:** Michael Knipe
**Liaison & Marketing:** Studio Blei, Milan
**Design:** Hollytree Publications Limited
**Illustrations:** Gerry Greaves


## Technology
A strong commitment to the advanced technological development of products and processes, with 1.700 research and development specialists working in 6 R & D centres in Brazil, France, Germany, Italy, the United Kingdom and the United States.

## Professionalism
The highly developed skills of its 62.000 employees, the modern industrial relations schemes, continuous training and retraining of personnel and labour organization systems at factory level, represent advanced management of human resources.

## Internationalism
One of the company's basic policies since its foundation has been to advocate community of interests between nations. Today, Pirelli operates 110 factories in 16 countries in Europe, North and South America, Africa and Australia.

## Diversification
A wide range of products and services, from telecommunication and energy transmission cables to tyres for most applications, from motor vehicle and industrial components to consumer products.

PIRELLI
**Ready for the future.**

Agenzia Centro

mass in the planet determines the size of the equatorial bulge: the slower the rotation and the more concentrated the mass, the smaller the bulge. Uranus has a rather large bulge. Since its rotation rate, about 17.2 hours, is lower than expected, the mass distribution may be less concentrated toward the center than the models predicted. Specifically, the heavy and light volatile substances may be mixed in a single layer, which would be more like a hot, dense atmosphere than an ocean.

Near the top of the atmosphere, where the temperature declines rapidly to about −221 degrees Celsius (it is as much as 10,000 degrees at the center of Uranus), the volatile subtances would presumably condense to form clouds. The top layer would be methane clouds, which *Voyager 2* did indeed observe. Moreover, in monitoring the top of the atmosphere the spacecraft discovered another baffling anomaly. Although Uranus now keeps its south pole directed toward the sun—the southern summer lasts for 21 years—the temperature there is no higher than it is above the equator or even above the north pole, which has been in darkness for many years. Clearly winds must be carrying heat away from the south pole, but the nature of the circulation is still unknown.

*Voyager 2*'s attentions were by no means confined to Uranus itself. As the spacecraft hurtled through the Uranian system it discovered two new rings, bringing the total to 11, and 10 tiny moons orbiting close to the planet. Yet the most sensational pictures were of the five outer, previously known moons. (Beginning with the innermost they are Miranda, Ariel, Umbriel, Titania and Oberon.) Seen from the earth through large telescopes the outer moons are mere points of light. The largest, Titania, is about 1,600 kilometers across, less than half the size of the earth's moon. All five are colorless: their fluffy, patchy surfaces consist primarily of water ice and an unidentified black carbonaceous material.

The Uranian moons present a strong contrast to those of Saturn. *Voyager 2* data indicate the Uranian moons are denser. They seem to be roughly half water ice and half rock, whereas most of the small moons of Saturn are thought to be 60 to 70 percent water ice and 30 to 40 percent rock. This discovery, although not entirely unexpected, contradicts simple models of how the solar system formed. Such models predict that one should find less rock and more ice as one moves farther from the sun.

The rockiness of the Uranian moons helps to explain the startling extent to which they seem to have been geologically active, albeit billions of years ago. Tectonic faulting or volcanism requires that the interior of the moon be hot enough to melt its ice. A principal source of internal heat is the decay of radioactive elements embedded in rock. The more rock in a moon, the stronger its radioactive heat source; but the smaller a moon is, the sooner it loses the heat it starts with to space. Before it was known how rocky the Uranian moons are they were widely thought to be too small to support much geologic activity. Although *Voyager 2* had found signs of activity on the moons of Saturn, this was attributed to frictional heating by tidal forces, which are considered insignificant on all the Uranian moons except Ariel.

Ariel does seem to have been the most active of the five moons. Its surface is extensively faulted, and in one region material has welled up from the interior through fissures and flowed out over craters and valleys. Yet all the moons have some features that require strong internal heating. The oddest features were found where they were least expected, on Miranda, the smallest moon: a V-shaped pattern of grooves called the chevron, two almost oval grooved patterns that resemble racetracks and a cliff some 20 kilometers high.

Voyager investigators have put forward a tentative explanation for Miranda's strangeness. The hypothesis assumes the moon was blasted apart by a large comet some time after it had differentiated, that is, after radiogenic heat had caused rock to settle to the core and ice to form an outer shell. As fragments of the shattered moon were drawn back together by their gravity, some rock would have ended up on the outside and some ice on the inside. Large blocks of rock would then have begun settling back to the core. The resulting stress at the surface might have produced expanding fracture patterns, at first jagged like the chevron but with time becoming rounded like the racetracks. Miranda's internal heat would probably not have sufficed to complete the process of differentiation a second time. Like the other Uranian moons it is now cold and geologically dead.

## *Cosmic Chronometer*

A group of astrophysicists has discovered a celestial object that emits an average of 186 highly regular bursts of radio waves every second. The interval between bursts varies cyclically, with a period of 12.3 days. By virtue of this combination of two rhythmic patterns the object, designated PSR 1855 + 09, has been identified as the fastest binary pulsar that is known.

A pulsar is thought to be a rapidly spinning neutron star. Such stars can arise from the gravitational collapse of a supernova's core. It is in conserving angular momentum as it shrinks to a diameter of only several kilometers that the neutron star attains its high rotational velocity. If the neutron star continuously emits a beam of electromagnetic radiation from a spot in the magnetized plasma overlying its surface, the beam is swept around like the beacon of a lighthouse. Such a radio beam, striking the earth with each revolution of the neutron star, can account for the observed radio-frequency pulsations.

When a pulsar happens to be one of a pair of stars orbiting each other, it is called a binary pulsar. Such a configuration can be detected (as in the case of PSR 1855 + 09) by a cyclical variation in the pulsar's signal, the result of a periodic Doppler shifting of the radio-burst frequency as the neutron star moves in its orbit toward and then away from the earth.

The systematic scan of the sky that turned up PSR 1855 + 09 was made with the National Astronomy and Ionosphere Center's 305-meter radio telescope near Arecibo in Puerto Rico by a graduate student, David Segelstein, under the direction of Joseph H. Taylor at Princeton University. Data confirming the finding were collected by Daniel Stinebring and Lloyd Rawley of Princeton and Aleksander Wolszczan of the Arecibo Observatory.

Although more than 400 pulsars have been catalogued, only five of these are known to have an orbiting companion star; PSR 1855 + 09 will be the sixth such object. Binary pulsars are of particular interest because from their observed orbital periods astrophysicists can estimate the mass of neutron stars.

Of even greater interest is the fact that the signals emanating from pulsars are often distinct and regular enough to serve as a natural clock of extremely high precision. Indeed, the fastest pulsar known, PSR 1937 + 21, has already proved to be comparable in accuracy and reliability to the best manmade atomic clocks; the same may be true for PSR 1855 + 09. Because the steady "ticking" of pulsars such as PSR 1937 + 21 and PSR 1855 + 09 is detected across vast galactic distances, it can be used to verify the existence of gravitational waves that may have been generated when the universe was less than a second old. The waves are deformations of space-time that ripple through the universe and could manifest themselves as unpredictable but correlated changes in the timing of the pulses from PSR 1937 + 21 and PSR 1855 + 09.

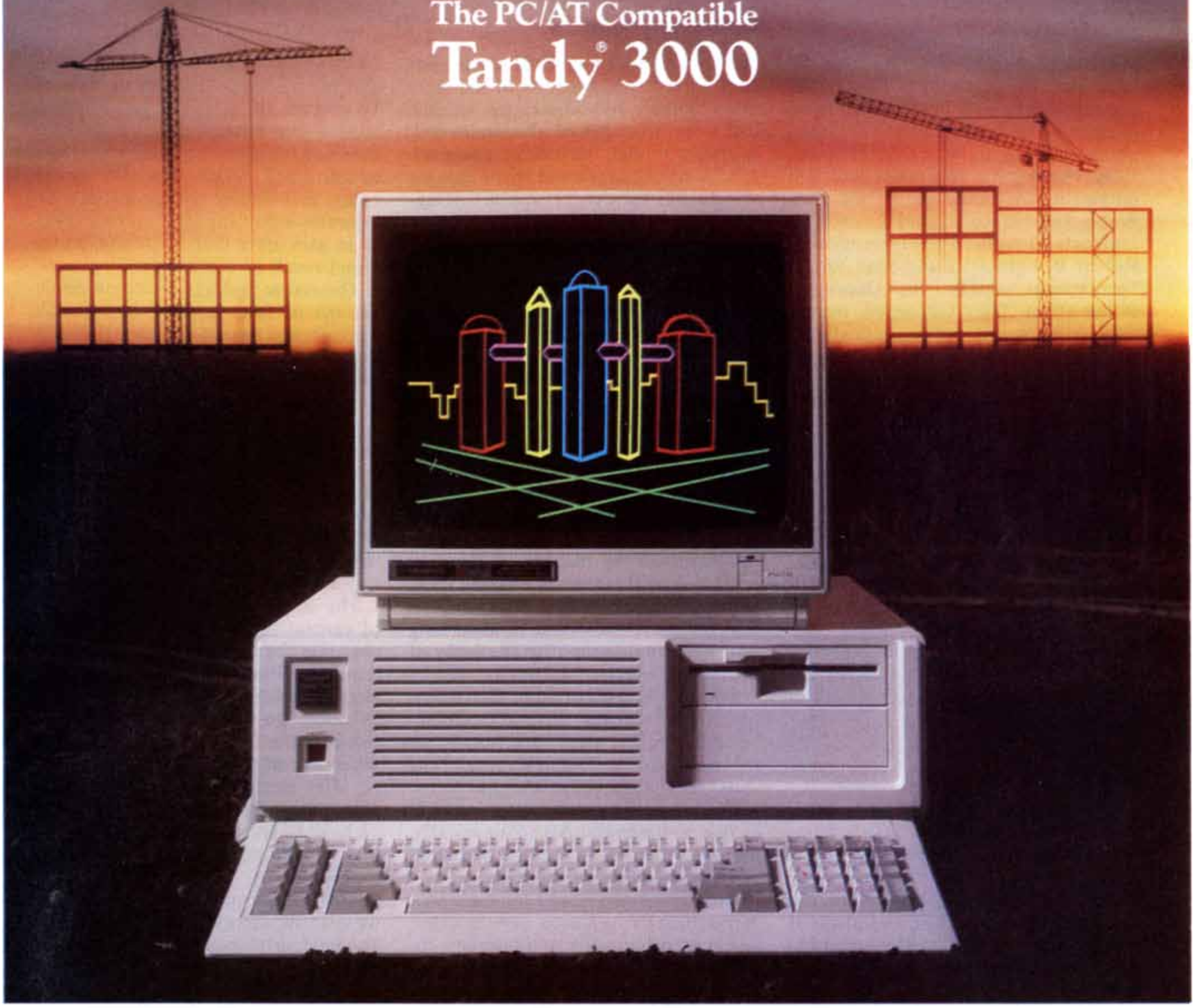
### *New Wave*

Light can exhibit particlelike properties and particles can exhibit wavelike properties. Although this duality has been verified repeatedly at the level of the atom, it is difficult to detect on a larger scale; for example, the wavelength of a running dog is less than the radius of a proton. Now Richard A. Webb and his colleagues at the IBM Thomas J. Watson Research Center have observed this fundamental physical reality on a superatomic scale in ordinary matter. They report their work in *Physical Review Letters.*

The IBM team exploited a variant of the Aharonov-Bohm effect, a phenomenon that was first predicted in 1959 by the theoretical physicists Yakir Aharonov of Yeshiva University and David Bohm of the University of London. In one demonstration of the effect a beam of electrons traveling through a vacuum is split in two by a negatively charged wire. The two partial beams pass around opposite sides of a solenoid (a cylindrical coil of wire) within which a magnetic field has been confined. The beams are subsequently recombined.

Aharonov and Bohm correctly predicted that if the two beams interact to form an interference pattern, the pattern will vary with the strength of the magnetic field in the solenoid. The reason is that the interference pattern depends on a property of the electron known as its phase, which is associated with its wavelike nature. The phase of an electron in turn can be shifted by a quantity known as the magnetic vector potential, which is associated with the magnetic field. The two theoreticians also noted that the physical size of the apparatus must be extremely small, since the maximum separation over which the two parts of the split beam can remain coherent, or in phase with respect to their initial values, is 60 micrometers.

The team at IBM decided to test whether the effect also occurs when electrons are traveling through ordinary materials. Would collisions between the electrons and the atoms in the material destroy the phase of the electrons? The workers fabricated minuscule gold rings ranging from .2 to .8 micrometer in diameter and from .04 to .13 micrometer in width. They immersed such a ring in a magnetic field and ran an electric current into it through one lead and out of it through another lead attached to the opposite side. They thereby split the current




# Making your ideas for today take shape for tomorrow.

To turn your ideas into reality, you need a tool for productive thinking: the Tandy® 3000.

The Tandy 3000 is compatible with programs designed for both the IBM® PC/AT and PC/XT. So whether your application involves statistical analysis, program development, database management or computer-aided design, you can get the job done on a Tandy 3000.

## We went beyond PC/AT compatibility

Because the Tandy 3000 operates at 8 megahertz, you get the job done 70% faster than most pc's, and 33% faster than the PC/AT. The Tandy 3000 handles complex operations with incredible efficiency.

## Power to share

The Tandy 3000 goes beyond conventional thinking, too. Since it is designed to use XENIX® as well as MS-DOS®, you can add low-cost display terminals so several associates can use the 3000 simultaneously.

## Tandy . . . Clearly Superior™

Take command of your future with the Tandy 3000—only $2599. Or choose the Tandy 3000 HD, with a 20-megabyte hard disk drive for just $3599.

***Available at over 1200 Radio Shack Computer Centers and at participating Radio Shack stores and dealers.***

**Radio Shack®**
**COMPUTER CENTERS**
A DIVISION OF TANDY CORPORATION

**Send me a Tandy 3000 brochure.**
Radio Shack, Dept. 86-A-571
300 One Tandy Center, Fort Worth, TX 76102

Name ____________________
Company ____________________
Address ____________________
City ____________________
State __________ ZIP __________
Phone ____________________

Prices apply at Radio Shack Computer Centers and participating stores and dealers. Monitor and MS-DOS not included. IBM/Registered TM International Business Machines Corp. MS-DOS and XENIX/Registered TM Microsoft Corp.

into two parallel paths, mimicking the Aharonov-Bohm configuration in the vacuum. To reduce interference from the motion of the gold atoms the team chilled the ring to near absolute zero.

By varying the magnetic field and measuring the resistance of the ring, Webb and his colleagues Corwin P. Umbach, Christian Van Haesendonck, Robert B. Laibowitz and Sean Washburn sought to detect the Aharonov-Bohm effect. They succeeded: the resistance rose and fell periodically as the magnetic field increased. Apparently the electrons in each side of the ring "remembered" their initial phase, so that their waveforms reinforced and interfered with one another.

The observance of the effect in ordinary materials may have important technological ramifications. Supriyo Datta of Purdue University notes, for instance, that the wavelengths of electrons in semiconductors are much longer than their corresponding wavelengths in metals. As a result the Aharonov-Bohm effect might be observed in relatively large semiconductor devices and might even be employed in their application.

## *Flights of Conjecture*

Whereas the wings of birds originated from limbs, insect wings had no such precursors. They must have evolved gradually from smaller winglike appendages. What is the use of such a structure? The evolution of flying insects during the Devonian period, some 400 to 350 million years ago, turns on the answer; only if such winglets had some adaptive value would they have evolved and enlarged into wings. Some investigators propose that the winglets had aerodynamic value from the start, enabling an insect to glide, float parachutelike on the wind or stabilize itself while falling.

Like Devonian dry-fly fishermen fashioning lures, Joel G. Kingsolver of Brown University and Mimi A. R. Koehl of the University of California at Berkeley examined such hypotheses by building and testing physical models of ancient insects. Their results, reported in *Evolution,* indicate that aerodynamics did not provide the initial driving force for the evolution of winglets. The work supports an alternative hypothesis: that the first winglike appendages acted as solar collectors, helping basking insects to raise their body temperature and become active.

The investigators formed the bodies of the test models, made in a variety of sizes and shapes, out of epoxy, a material that is similar to the tissue of present-day insects in density and thermal conductivity. For aerodynamic tests the workers added wings of plastic film on a wire frame; for tests of the thermoregulatory hypothesis the models had wings of paper or aluminum foil. Wingless models served for comparison. A wind tunnel and strain gauges attached to the models enabled Kingsolver and Koehl to measure aerodynamic forces, and a heat lamp and thermocouples revealed the effect of wings on body temperature.

The investigators found that wings shorter than a centimeter or so had no effect on the aerodynamic characteristics of the model, regardless of body size or shape. They did not produce the lift needed for gliding or add to the drag that enables an insect to travel great distances on the wind; nor did they exert stabilizing forces on a model placed askew in the air current. Apparently vortexes generated at the wingtips dissipate the meager lift or drag produced by short wings, making them aerodynamically useless. The winglets were effective, however, in absorbing heat and conducting it to the body of the model insects.

For wing lengths of more than a centimeter or two the findings were reversed. Longer wings did not warm the models more effectively than short wings; heat absorbed farther out on the wings seems to be transmitted less efficiently to the body. Aerodynamically, however, the wings became increasingly effective as they lengthened.

Kingsolver and Koehl conclude that aerodynamics replaced thermoregulation in driving the evolutionary increase in wing length once the winglets reached a length of about a centimeter and began generating effective lift or drag. Because the wing length at which the transition occurs is less in relation to body size for larger insects, the authors suggest that a simple increase in size may have been the event that pointed insects down the evolutionary road to flight. A proportionate enlargement of wings and body could at a stroke have changed the appendages from solar collectors to airfoils.

## *Weighty Problem*

People who want to quit smoking are frequently deterred by the specter of gaining weight. Their fears are reinforced by statistics showing that nonsmokers are generally heavier than smokers. It is often assumed that such differences are mediated by appetite. Smoking, the theory goes, gratifies oral needs. After quitting, food is substituted for cigarettes as a source of oral gratification. The result is weight gain.

Recently, however, a group of investigators at the University of Lausanne in Switzerland has found that smoking 24 cigarettes a day increases the energy expenditure of the human body by 10 percent. If a person gives up cigarettes, the energy that would have been consumed by smoking is converted into body weight. The authors calculate that with no change in diet a 24-cigarette-a-day smoker can be expected to gain more than 20 pounds as the direct result of abstinence.

The recent findings, based on observations of four healthy young male smokers and four of their female counterparts, were published in the *New England Journal of Medicine.* Each subject was observed twice, once while smoking 24 cigarettes in 24 hours and once while abstaining for an equal period. Both 24-hour periods were spent in an experimental chamber equipped with devices for measuring levels of oxygen and carbon dioxide and a radar system for measuring the subject's physical activity.

The measurements of oxygen and of carbon dioxide were employed to compute the energy consumed in metabolism. When not smoking, the subjects expended 2,230 kilocalories in 24 hours on the average. When they smoked, the figure rose to 2,445. The difference of about 10 percent persisted after controlling for the subjects' activities in the chamber. The elevated level of energy consumption lasted through the night but disappeared by morning.

The precise mechanism underlying the rise in energy expenditure is not known. One clue comes from the finding that when the subjects smoked, the heartbeat was 20 percent faster and 45 percent more noradrenaline was excreted. Noradrenaline has an important role in the sympathetic nervous system, which regulates the action of the heart and the involuntary muscles. The authors suggest that nicotine may arouse the sympathetic nervous system, which in turn stimulates the activity of the heart and other internal organs.

## *Taking the Waters*

Bath, Baden Baden, Saratoga Springs and many other old spas testify to the great age of the notion that immersion in warm mineral water is good for what ails you. Yet little work has been done over the centuries to find out what the precise physiological effects of the treatment are. Eight workers at the British Royal Infirmary in Bristol have now looked into the matter. They report their results in the *British Medical Journal.*

First the Bristol group set up its own spa by bringing in 500 gallons of water from Bath. The investigators filled

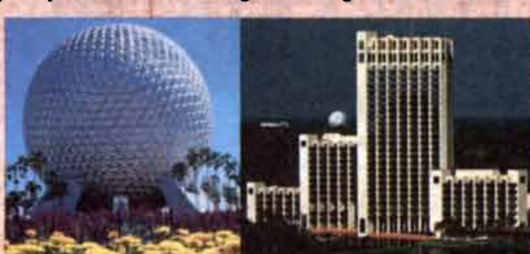
# GMC TRUCK GREAT ESCAPE SWEEPSTAKES

# At last, a sweepstakes for individuals who'd never dream of entering one.

Many contests try to give away things you probably wouldn't want at <u>any</u> price. But our GMC Truck Great Escape Sweepstakes offers something you can't put a price on: Escape. It's a chance to win the freedom to get away and experience the great American outdoors. And the opportunity to take off in a brand-new GMC truck. Even if you haven't entered the more mundane sorts of sweepstakes before, you should consider the unusually attractive opportunities presented by the GMC Truck Great Escape Sweepstakes:

**GRAND PRIZE: A 1986 GMC SAFARI VAN, YOUR CHOICE OF A REGAL SEBRING 195XL BOWRIDER BOAT OR A 1986 STARCRAFT 20-FT TRAVEL TRAILER, PLUS A GREAT ESCAPE FOR FOUR.*** You could win a stylish and accommodating 1986 GMC Safari with the towing capacity no front-wheel-drive mini-van can match, plus your choice of a Regal Sebring 195XL bowrider boat or a 1986 Starcraft 20-ft travel trailer, plus a ten-day, nine-night trip for four to the Grand Canyon, Yellowstone and Yosemite National Parks. Air fare, lodging, $1,000 spending money, and use of a vehicle at each site included.

**FIRST PRIZE: 1986 GMC SAFARI VAN, PLUS YOUR CHOICE OF A REGAL SEBRING 195XL BOWRIDER BOAT OR A 1986 STARCRAFT 20-FT TRAVEL TRAILER.***



©1986 Walt Disney Productions.

**TWO SECOND PRIZES: TRIP FOR FOUR TO THE WALT DISNEY WORLD® RESORT.***
Four days, three nights of sightseeing, adventuring, relaxing. Includes lodging at the Buena Vista Palace, inside the Walt Disney World Village, and three-day passports to the Magic Kingdom® and EPCOT® Center.

**PLUS TEN THIRD PRIZES OF COLEMAN CAMPING OUTFITS AND THOUSANDS OF INSTANT WINNER PRIZES, TOO!**

**FOUR BONUS PRIZES FOR DEALERSHIP ENTRIES <u>ONLY</u>: A 1986 GMC LIGHT-DUTY TRUCK, PLUS YOUR CHOICE OF A STARCRAFT BOAT OR A 1986 STARCRAFT CAMPER.***
Only those who register at their GMC Truck dealer are eligible for this great bonus: a choice of a 1986 GMC Safari, S-15 Pickup, S-15 Jimmy or Full-Size Pickup, <u>plus</u> a choice of a 1986 Starcraft 24-ft Galaxy camping trailer or a Starcraft 16-ft Medalist boat and trailer.

*Prizes may vary from illustrations shown due to model changes, availability, and other circumstances. See your participating GMC Truck dealer for entry forms, official rules and details. No purchase necessary. Must be 18 or older to enter. Enter at dealer by May 31, 1986. Void where prohibited.

Let's get it together . . . buckle up. GM a tank with the water and maintained the temperature at 35 degrees Celsius (slightly less than body temperature). Four men and four women, all apparently in good health, sat chest-deep in the water for two hours.

According to the investigators, in tests following immersion the subjects showed "highly significant diuresis" (increased urine flow), a twofold increase in the excretion of sodium, a significant increase in the excretion of potassium, a 5 percent decline in the indexes relating to red blood cells (hemoglobin, packed cell volume and red-cell count), a small fall in the viscosity of the blood plasma, a 50 percent increase in cardiac output and a loss in weight that averaged half a kilogram.

Does all of this do anything for the bather's health? The Bristol group noted that the increased cardiac output was not accompanied by a change in blood pressure, meaning "there must have been a corresponding fall in peripheral resistance" and a rise in peripheral blood flow, which "may have beneficial effects" in some diseases. The investigators also noted a reduction in edema, or accumulation of fluid in the body. They conclude that the therapeutic effects of taking the waters merit further study.

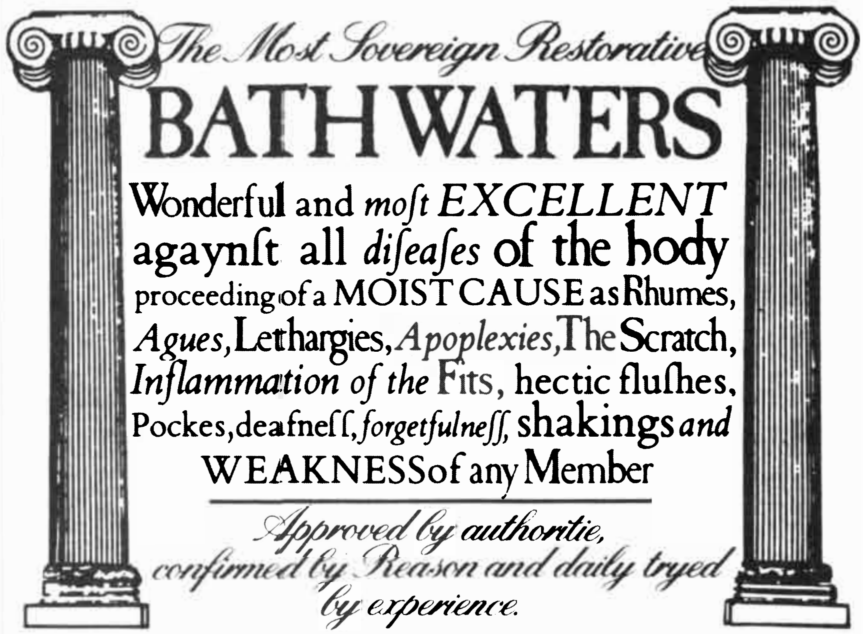
Such effects, if any, can be had at considerably less cost than the price of a trip to Bath, Baden Baden or Saratoga Springs. The Bristol group reports that two hours in a tub filled with tap water maintained at 35 degrees C. produces essentially identical results.

## *Three R's*

For the foreseeable future, biological and behavioral investigators and (to a lesser extent) product testers will continue to require live animals as subjects, but they can attempt to reduce the number of animals that are involved and can minimize the subjects' pain and distress. So says the Congressional Office of Technology Assessment (OTA) in a report titled *Alternatives to Animal Use in Research, Testing, and Education.*

The document acknowledges that there has been increasing public demand for the development of alternatives to animal experimentation as it is currently carried out. It describes the feasible alternatives as "the three R's": replacement with nonanimals, reduction in the numbers of animals required and refinement of techniques in order to reduce their pain and distress.

Among the possible replacements for animals are cultures of organs, tissues and cells, and chemical and physical systems that mimic biological systems. (For instance, a new approach to testing for eye irritancy exposes a membrane of the chick embryo to a potential irritant and gauges the severity of the reaction by the degree of damage to the membrane.) Another option, computer simulation, is also promising and is being widely exploited in education. Even when computers can be resorted to for research and testing, however, the report indicates that some animals will still be needed to provide the data on which the soft-




**VIRTUES OF SPA were extolled in an 18th-century card from Bath. Recent experiments suggest that immersion in Bath water and immersion in tap water have the same effects.**

# WHAT ARE ZERO COUPON CERTIFICATES AND WHY DO THEY BELONG IN MY IRA?

**Think of Zeros as $1,000 bills at half-price.**

That's the easiest way to describe the newest IRA option from Wells Fargo Bank. Each Zero Coupon Certificate has a guaranteed maturity value of $1,000. But costs just a fraction of that amount–as little as $500. That's half-price!


Open your IRA now with tax-sheltered Zeros at Wells Fargo Bank. Unlike many investments, Zeros can only go one way–up.*

**Your investment is guaranteed. And federally insured.**

Now you can plan ahead with confidence. Choose the exact maturity date of each Zero. And know exactly what each will be worth if held to maturity. We guarantee it. In addition, Wells Fargo's Zero Coupon Certificates are insured by the F.D.I.C.

**The longer you hold onto your Zeros, the less you pay now.**

Decide when you want your Zeros to mature. Then read across on the chart below to discover how little you pay compared to the face value.

| Time To Maturity | Value At Maturity | Price Per Zero |
|---|---|---|
| 5 years | $1,000 | $657 |
| 6 years | $1,000 | $597 |
| 7 years | $1,000 | $545 |
| 7+ years | $1,000 | $500 |

*All figures and statements assume Zero Coupon Certificates will be held to maturity. Chart shows example of prices based on a daily compounded rate of 8.402% with a yield of 8.764% for the 5-year term to a daily compounded rate of 8.770% with a yield of 9.165% for the 7+ year term.

Penalty for early withdrawal.

**A "Half-Price" Zero brings twice the rewards.**

For just $500, you can buy a unique Wells Fargo Zero Coupon Certificate that will double in value at its maturity. Begin adding some Zeros to your IRA each year while you're still young enough, and by the time you retire you may be able to pay off a mortgage or cruise the world. Or both.

**Use Zeros to protect against risk.**

J. P. Morgan said that the stock market would go up and the market would go down. By diversifying your IRA portfolio with Wells Fargo Zero Coupon Certificates, you won't be caught with your stocks down when you need the money the most.

**Wells Fargo guarantees to redeem your Zeros before maturity.**

If necessary, you can redeem your Zeros early. Based on the interest climate at that time, you could receive more or less than your original investment.

**Zeros offer an excellent IRA rollover opportunity.**

It may be advantageous to roll over contributions from another IRA or qualified plan into Wells Fargo's IRA Zero Coupon Certificates. For example, if you and your working spouse have $10,000 in IRA funds, you could purchase up to 20 Zeros and another 8 with this year's IRA contribution. For a total guaranteed value at maturity of $28,000.

**You don't need a broker to buy Zero Coupon Certificates.**

With over one billion dollars in IRA assets alone, Wells Fargo is the choice of Californians who want growth and stability. And a broad range of IRA investment options. Let our IRA Specialists help you make the most of your money. Visit a Wells Fargo office or call us toll free for current rates.

**CALL 1-800-BEST IRA**

**WELLS FARGO BANK**

Member F.D.I.C.

ware is based and to prove the validity and reliability of any program.

Methods for reducing the number of animal subjects are also varied. Studies can be designed to yield reliable information with fewer animals; research projects can be better coordinated, for example to allow workers to remove several tissues from an animal at one time. Refinements suggested by the OTA include administration of more precise amounts and types of anesthetics, more careful monitoring of pain, the use of instruments that are noninvasive or that produce minimal pain and the humane destruction of animals before they recover from surgical anesthesia.

In addition to describing alternatives to animal use, the OTA report gives Congress a range of steps it might take to encourage investigators, testers and educators to adopt alternatives. Congress could, for instance, go so far as to ban either all animal experimentation or the use of certain animals or procedures; it could fund studies to develop alternatives, or it could mandate the creation of new data bases to avoid unintentional duplication of experiments. The OTA stops short of recommending one action over another, although it does state that live animals will continue to be required for observing the complex interactions of cells, tissues and organs and that any outright ban on the use of animals in research, testing and education "would effectively arrest" most basic biomedical and behavioral research and toxicology testing in the U.S. One example of the kind of work that would be halted by such a ban is a series of animal trials now under way to test the efficacy of a potential vaccine against AIDS.

## *The MCC Experiment*

The Microelectronics and Computer Technology Corporation (MCC) of Austin, Tex., was conceived in 1982 to address two challenges, one technical and the other economic. Four years later the company remains an experiment as exciting—and chancy—as any conducted in its elaborately equipped facilities.

The technical challenge arises because current methods of building and operating computers have progressed to the point where fundamental physical limitations hinder further significant gains in speed, power and miniaturization. The next generation of technology will most probably be based, at least in part, on new principles. Inventing the new technology will involve a good deal of risk: investigators are bound to take many wrong turns, and any successes are not likely to show financial profits for some time to come. The research that is needed is not the kind that would naturally be carried out in universities; yet most individual corporations, which must operate with an eye on their short-term performance, cannot invest the necessary time, money and talent to create a completely new technology. Lending urgency to this problem, from the point of view of American high-technology companies, is competition from Japan, Inc. The orchestrated and cooperative relations among Japanese industry, finance, science, labor and government make Japanese enterprises able to compete formidably in any electronics market that is not protected, or privileged by national security.

In early 1982 William C. Norris, chairman of the Control Data Corporation, believed he had part of the answer to these problems. Norris' idea was to combine the technical and financial resources of a large number of American corporations to form a new kind of research institution: the private-sector research cooperative. Norris' idea has a Japanese counterpart in the interdisciplinary research institution set up to develop technology for the so-called fifth generation of computing technology. In the U.S. Norris' idea was realized as MCC.

An extraordinary concept, MCC was also unusual in that it found widespread bipartisan support on Capitol Hill. Congress removed a major restraint to the founding of MCC by passing the Cooperative Research Act of 1984, which allowed competitors to work together on certain kinds of long-range research projects without fear of antitrust regulations. Members of Congress of varied political persuasion were friendly to the MCC concept because, as William D. Stotesbery, MCC's director of government and public affairs, puts it, "we were in the fortunate position of having liberals view us as an extension of industrial policy and conservatives as an extension of laissez faire."

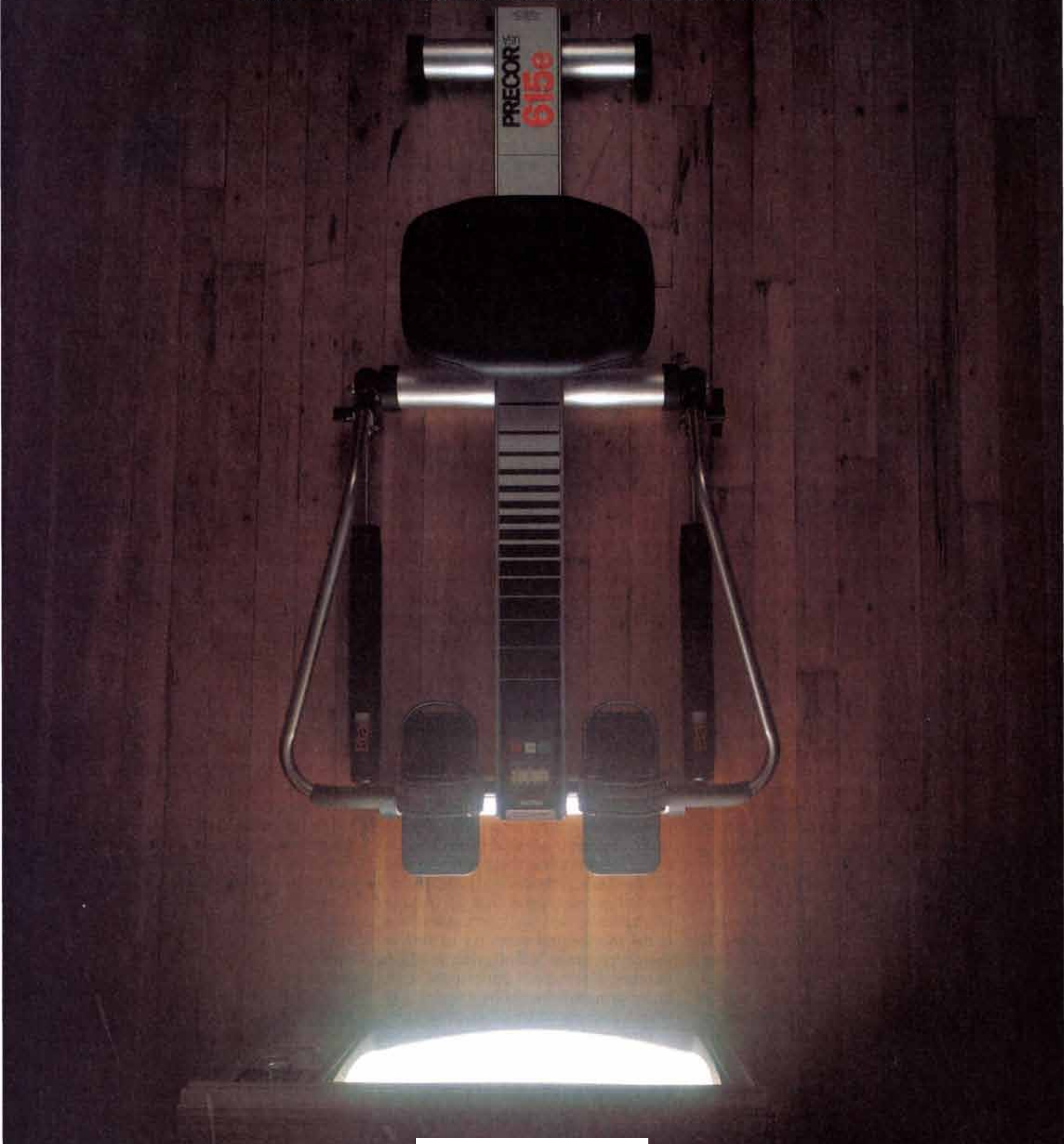
Industry's response was also favorable. When MCC was founded, it had 12 member companies. It has since grown to 21: Advanced Micro Devices, Allied, Bellcore, BMC, Boeing, Control Data, Digital Equipment, Eastman Kodak, Gould, Harris, Honeywell, Lockheed, Martin Marietta, Minnesota Mining and Manufacturing, Mostek, Motorola, National Semiconductor, National Cash Register, RCA, Rockwell and Sperry.

For the price of a share ($150,000 when MCC was founded, $500,000 when Bellcore, its most recent member, joined in 1985) each company receives the right to spend more mon-

The most efficient aerobics program on TV is sitting right in front of it. On a Precor 615e rower. It gives you a total body workout in less than half of *60 Minutes*. With silent, smooth, solid state design. And digital electronics that monitor every stroke. Channel your energy. Call 1-800-551-7722, for the Precor dealer nearest you. And fine-tune your body. While you watch.

**PRECOR USA**
Personal Power Tools

# STAY TUNED FOR FURTHER DEVELOPMENTS.



ey: to invest in one or more of the cooperative's four research programs. In joining MCC a company buys what is essentially a highly leveraged research program: for each dollar invested, the company is allowed to participate in many dollars' worth of research. On joining, a company makes a commitment to stay in a research program for at least three years; after that time it must give one year's notice of any intention to withdraw.

The object of MCC's research is to raise the general level of technology, not to produce specific products. MCC's deliverables are not chips or machines but rather techniques, configurations and approaches that the "customers"—the shareholder corporations—can apply to the development of new products or can incorporate into products. In many cases MCC will deliver not only the systems themselves but also the methods by which the systems were developed.

Companies take part in research programs in two ways: by direct financial support and by assigning personnel to join MCC's research staff. Some of the assigned employees split their time between doing pure research and maintaining liaison between MCC and the parent company; the rest of the employees assigned by parent companies have the same status as the investigators employed directly by MCC. Of MCC's 340-person research staff, 17 percent are liaison staff, 16 percent were assigned to MCC programs by their parent companies and 67 percent were hired directly by MCC. Finding staff was difficult at first. Retired admiral Bobby R. Inman, MCC's president, suggests that "companies were not always willing to offer their best talent, particularly if that talent was working on near-term projects." MCC hired many more research staff from outside parent companies than it had originally intended.

The four research programs are "packaging" (developing new ways to connect the elements on a microchip with one another and with the machinery in which the chip resides); "software technology" (producing procedures and computerized tools that simplify the task of writing complex programs for specialized applications); "very large scale integration/computer-aided design [VLSI/CAD]" (developing a similar set of tools that can be applied to the design of complex integrated circuits), and "advanced computer architectures." The last program is itself split into four separate research programs: "parallel processing" (hardware and software that enable computers to operate at high speed and to solve certain kinds of complex problems by performing many parts of a task simultaneously); "data-base-system management" (methods of storing and retrieving complex information); "human-factors technology" (flexible means of communication between human beings and computers, perhaps including speech or handwriting), and "artificial intelligence [AI]/knowledge-based systems" (computer systems that make decisions by applying new ways of storing knowledge).

Because of MCC's special nature, its projects differ in significant ways from those undertaken in academia or industry. One difference is that they may not pay off for many years. One of the projects in the AI/knowledge-based systems program, for example, is to find a way of storing and representing the "commonsense" knowledge that human beings bring to decision making. To amass commonsense knowledge, MCC is hiring college students to read encyclopedia articles and to identify the ideas or concepts each article assumes the reader knows. Representing such knowledge in a computer memory is a trickier problem, and Douglas B. Lenat, who heads this research team, estimates that gathering the knowledge base alone will take between 200 and 300 man-years.

Another characteristic of MCC's programs is how much money they have at their disposal. John R. Hanne, program director of VLSI/CAD, notes that "no public company could do what we've done, because what we've done is to overcapitalize," investing in a high ratio of expensive equipment to researchers. On the other hand, he adds, "no university could put together electrical engineers, computer scientists, physicists and mathematicians—all of whom are necessary for this work—in a stable situation, because the standard unit in the university is the department." Another advantage is that "I get data, advice and feedback from the 11 different companies participating in the VLSI/CAD program, rather than just one company."

The unusual structure of MCC leads to some unconventional difficulties as well. The most notable difficulty concerns cooperation between research programs. In a sense each program is a limited partnership with a different set of partners, and so information that is the intellectual property of one research program is not given freely to another; if it were, a participant in the latter program would benefit from research sponsored by other companies.

This problem is particularly troubling because the research programs have much to gain from one another. For example, John T. Pinkston, MCC's chief scientist, notes that the process of designing specialized software and that of designing a specialized microchip have much in common. In both procedures one must gradually reduce what may be an ill-defined "behavioral" goal to a specific set of procedures and mechanisms. Hence "although the software technology and VLSI/CAD programs cannot quite trade software, they can still understand each other's approaches and problems" and would both profit from close cooperation.

MCC has addressed the problem legalistically: whenever two programs cooperate, they sign a contract, specifying who is to give what to whom. Every formal cooperation is documented by a written report. Several of the program directors have expressed a mild sense of frustration at the complications of doing the paperwork necessary for cooperation with other programs. MCC's executives say that the most important change they would make in the cooperative's structure if they were to start again from scratch would be to make every shareholder participate in every research program. They acknowledge that such a stipulation might initially have discouraged most if not all of the parent companies from joining, since many companies are interested in only one or two of MCC's projects.

Facilitating communication among research programs has been a major priority of Inman's since late in 1984, he says, about two years after he was appointed MCC's president. He believes the programs have much to gain from one another, but he adds that a formal system of documenting various programs' contributions to any joint effort is necessary, because "fairness and equitability are critical in maintaining the shareholders' trust." He is optimistic about the future of communication among programs: "The members now want more synergism; they are less concerned about leakage. They were suspicious at the outset, but they have been getting friendlier."

Inman's background in Government service—he is a former director of Naval Intelligence and of the National Security Agency and a former deputy director of the Central Intelligence Agency—has prepared him well for managing a research effort that brings together competing interests. "At NSA," Inman recalls, "the problem was, How do you get diverse agencies and services to work together when they don't want to? Before I joined MCC I thought, Do these companies know they must work together? The answer was yes; our Japanese colleagues provided that."

Part of Inman's approach is to help

the companies understand one another's priorities and needs. One of his techniques is to hold what he calls an annual recalibration of goals: "Every 12 to 18 months, in each program, I get all the participating shareholders together—without a representative from MCC's management—to discuss goals. Then in the afternoon I come in and they brief me. I want to get them to confront the reality that they have differing views about goals and process."

An even more important priority of Inman's has been technology transfer: ensuring that the shareholders are able to use the new technology developed at MCC. Inman notes that MCC's research effort "speeds up only part of the process; it does not by itself guarantee successful competition in the marketplace." MCC will not have been a success if its advances are not incorporated into the shareholders' products and methods of operation.

The move from Government service to the private sector has given Inman new perspectives on problems the U.S. faces in international competition. He says: "I have spent 30 years of my career looking at the outside world and the last three and a half looking at American companies, studying the emerging technology and our investment in education. I have moved from asking [when MCC was starting up] 'Where is the talent?' to asking 'How do you grow it?'... We must increase our investment in education at all levels, we must work away from specialization in high school and college, we must put in place a process for retraining, and we must emphasize continuing education in the workplace."

How well is MCC working? Can such a cooperative work at all? It is still too early to answer either question; most of MCC's projects will not bear fruit until a few years from now, and even then it will not be certain whether the new technology can find its way into products. It is clear, however, that cooperative research efforts such as MCC can be only part of the solution to America's weakening position in international competition. "I stand back from my MCC job," Inman muses, "and I ask: Where are the midsize and smaller firms going to get advice on manufacturing technology to compete in the international marketplace?... I think about the impact new technologies can have on increasing productivity—how to attract capital to that?" To compete successfully the U.S. will need more than interindustry cooperative research laboratories. It may, indeed, need to emulate that ultimate Japanese innovation, the pattern of coordination among the various elements of a country's economy.



# Islamic Astronomy

*While astronomy withered in medieval Europe it flourished in Islam. Renaissance astronomers learned from the texts of Islamic scholars who had preserved and transformed the science of the ancient Greeks*

by Owen Gingerich

Historians who track the development of astronomy from antiquity to the Renaissance sometimes refer to the time from the eighth through the 14th centuries as the Islamic period. During that interval most astronomical activity took place in the Middle East, North Africa and Moorish Spain. While Europe languished in the Dark Ages, the torch of ancient scholarship had passed into Muslim hands. Islamic scholars kept it alight, and from them it passed to Renaissance Europe.

Two circumstances fostered the growth of astronomy in Islamic lands. One was geographic proximity to the world of ancient learning, coupled with a tolerance for scholars of other creeds. In the ninth century most of the Greek scientific texts were translated into Arabic, including Ptolemy's *Syntaxis,* the apex of ancient astronomy. It was through these translations that the Greek works later became known in medieval Europe. (Indeed, the *Syntaxis* is still known primarily by its Arabic name, *Almagest,* meaning "the greatest.")

The second impetus came from Islamic religious observances, which presented a host of problems in mathematical astronomy, mostly related to timekeeping. In solving these problems the Islamic scholars went far beyond the Greek mathematical methods. These developments, notably in the field of trigonometry, provided the essential tools for the creation of Western Renaissance astronomy.

The traces of medieval Islamic astronomy are conspicuous even today. When an astronomer refers to the zenith, to azimuth or to algebra, or when he mentions the stars in the Summer Triangle—Vega, Altair, Deneb—he is using words of Arabic origin. Yet although the story of how Greek astronomy passed to the Arabs is comparatively well known, the history of its transformation by Islamic scholars and subsequent retransmission to the Latin West is only now being written. Thousands of manuscripts remain unexamined. Nevertheless, it is possible to offer at least a fragmentary sketch of the process.

## The House of Wisdom

The foundations of Islamic science in general and of astronomy in particular were laid two centuries after the emigration of the prophet Muhammad from Mecca to Medina in A.D. 622. This event, called the Hegira, marks the beginning of the Islamic calendar. The first centuries of Islam were characterized by a rapid and turbulent expansion. Not until the late second century and early third century of the Hegira era was there a sufficiently stable and cosmopolitan atmosphere in which the sciences could flourish. Then the new Abbasid dynasty, which had taken over the caliphate (the leadership of Islam) in 750 and founded Baghdad as the capital in 762, began to sponsor translations of Greek texts. In just a few decades the major scientific works of antiquity—including those of Galen, Aristotle, Euclid, Ptolemy, Archimedes and Apollonius—were translated into Arabic. The work was done by Christian and pagan scholars as well as by Muslims.

The most vigorous patron of this effort was Caliph al-Ma'mūn, who acceded to power in 813. Al-Ma'mūn founded an academy called the House of Wisdom and placed Ḥunayn ibn Isḥaq al-'Ibādī, a Nestorian Christian with an excellent command of Greek, in charge. Ḥunayn became the most celebrated of all translators of Greek texts. He produced Arabic versions of Plato, Aristotle and their commentators, and he translated the works of the three founders of Greek medicine, Hippocrates, Galen and Dioscorides.

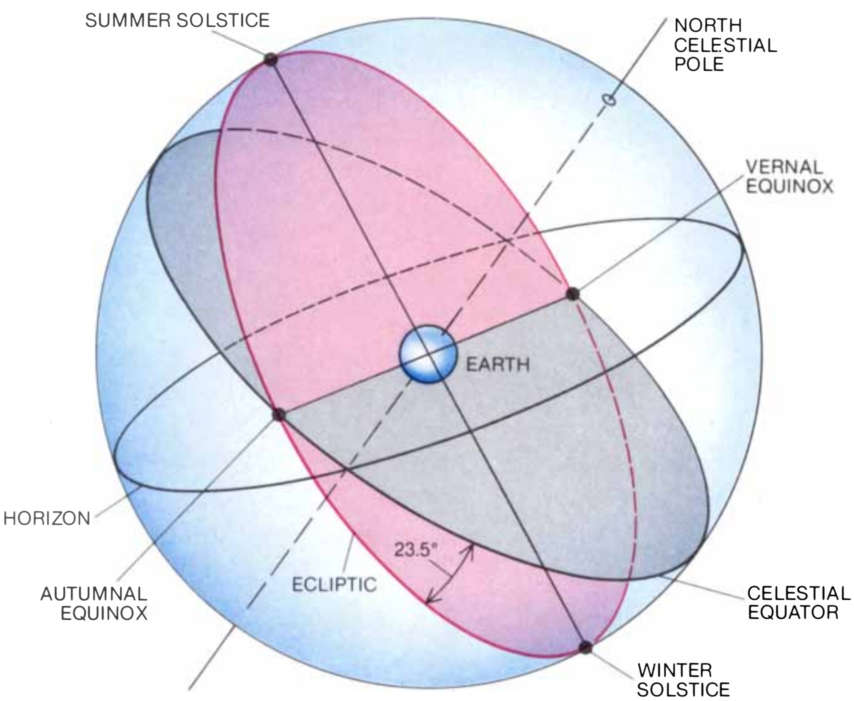
The academy's principal translator of mathematical and astronomical works was a pagan named Thābit ibn Qurra. Thābit was originally a money changer in the marketplace of Harran, a town in northern Mesopotamia that was the center of an astral cult. He stoutly maintained that the adherents of this cult had first farmed the land, built cities and ports and discovered science, but he was tolerated in the Islamic capital. There he wrote more than 100 scientific treatises, including a commentary on the *Almagest.*

Another mathematical astronomer at the House of Wisdom was al-Khwārīzmī, whose *Algebra,* dedicated to al-Ma'mūn, may well have been the first book on the topic in Arabic. Although it was not particularly impressive as a scientific achievement, it did help to introduce Hindu as well as Greek methods into the Islamic world. Sometime after 1100 it was translated into Latin by an Englishman, Robert of Chester, who had gone to Spain to study mathematics. The translation, beginning with the words "Dicit Algoritmi" (hence the modern word algorithm), had a powerful influence on medieval Western algebra.

Moreover, its influence is still felt in all mathematics and science: it marked the introduction into Europe of "Arabic numerals." Along with certain trigonometric procedures, the Arabs had borrowed from India a system of numbers that included the zero. The Indian numerals existed in two forms in the Islamic world, and it was the Western form that was transmitted through Spain into medieval Europe. These nu-

**STAR MAP of the constellation Perseus is a medieval Islamic copy of a drawing made in the 10th century by the Persian astronomer 'Abd al-Raḥmān al-Ṣūfī. Al-Ṣūfī revised the star catalogue compiled in the second century by Ptolemy, the foremost astronomer of ancient times. The illustration is a page of a manuscript that now belongs to the Egyptian National Library in Cairo.**


صورت برساوس چنانکه در کره است




**CELESTIAL SPHERE revolving about the earth was conceived by ancient astronomers to explain the motions of the sun and stars. Its poles and equator are projections onto the sky of the earth's poles and Equator. As the sphere turns daily on its axis from east to west the sun and stars move around the poles along concentric circles. The sun moves slightly slower than the stars, and so in the course of a year it follows an easterly, circular path, called the ecliptic, on the celestial sphere. At the vernal and autumnal equinoxes the sun crosses the celestial equator; at the summer and winter solstices it is about 23.5 degrees above or below the equator. An observer on the earth sees only the celestial hemisphere above the horizon.**

merals, with the explicit zero, are far more efficient than Roman numerals for making calculations.

Yet another astronomer in ninth-century Baghdad was Aḥmad al-Farghānī. His most important astronomical work was his *Jawāmi'*, or *Elements*, which helped to spread the more elementary and nonmathematical parts of Ptolemy's earth-centered astronomy. The *Elements* had a considerable influence in the West. It was twice translated into Latin in Toledo, once by John of Seville (Johannes Hispalensis) in the first half of the 12th century, and more completely by Gerard of Cremona a few decades later.

Gerard's translation of al-Farghānī provided Dante with his principal knowledge of Ptolemaic astronomy. (In the *Divine Comedy* the poet ascends through the spheres of the planets, which are centered on the earth.) It was John of Seville's earlier version, however, that became better known in the West. It served as the foundation for the *Sphere of Sacrobosco*, a still further watered-down account of spherical astronomy written in the early 13th century by John of Holywood (Johannes de Sacrobosco). In universities throughout Western Christendom the *Sphere of Sacrobosco* became a long-term best seller. In the age of printing it went through more than 200 editions before it was superseded by other textbooks in the early 17th century. With the exception of Euclid's *Elements* no scientific textbook can claim a longer period of supremacy.

Thus from the House of Wisdom in ancient Baghdad, with its congenial tolerance and its unique blending of cultures, there streamed not only an impressive sequence of translations of Greek scientific and philosophical works but also commentaries and original treatises. By A.D. 900 the foundation had been laid for the full flowering of an international science, with one language—Arabic—as its vehicle.

## Religious Impetus

A major impetus for the flowering of astronomy in Islam came from religious observances, which presented an assortment of problems in mathematical astronomy, specifically in spherical geometry.

At the time of Muhammad both Christians and Jews observed holy days, such as Easter and Passover, whose timing was determined by the phases of the moon. Both communities had confronted the fact that the approximately 29.5-day lunar months are not commensurable with the 365-day solar year: 12 lunar months add up to only 354 days. To solve the problem Christians and Jews had adopted a scheme based on a discovery made in about 430 B.C. by the Athenian astronomer Meton. In the 19-year Metonic cycle there were 12 years of 12 lunar months and seven years of 13 lunar months. The periodic insertion of a 13th month kept calendar dates in step with the seasons.

Apparently, however, not every jurisdiction followed the standard pattern; unscrupulous rulers occasionally added the 13th month when it suited their own interests. To Muhammad this was the work of the devil. In the Koran (chapter 9, verse 36) he decreed that "the number of months in the sight of God is 12 [in a year]—so ordained by Him the day He created the heavens and the earth; of them four are sacred: that is the straight usage." Caliph 'Umar I (634–44) interpreted this decree as requiring a strictly lunar calendar, which to this day is followed in most Islamic countries. Because the Hegira year is about 11 days shorter than the solar year, holidays such as Ramadan, the month of fasting, slowly cycle through the seasons, making their rounds in about 30 solar years.

Furthermore, Ramadan and the other Islamic months do not begin at the astronomical new moon, defined as the time when the moon has the same celestial longitude as the sun and is therefore invisible; instead they begin when the thin crescent moon is first sighted in the western evening sky. Predicting just when the crescent moon would become visible was a special challenge to Islamic mathematical astronomers. Although Ptolemy's theory of the complex lunar motion was tolerably accurate near the time of the new moon, it specified the moon's path only with respect to the ecliptic (the sun's path on the celestial sphere). To predict the first visibility of the moon it was necessary to describe its motion with respect to the horizon, and this problem demanded fairly sophisticated spherical geometry.

Two other religious customs presented problems requiring the application of spherical geometry. One problem, given the requirement for Muslims to pray toward Mecca and to orient their mosques in that direction, was to determine the direction of the holy city from a given location. Another problem was to determine from celestial bodies the proper times for the prayers at sunrise, at midday, in the afternoon, at sunset and in the evening.

Solving any of these problems involves finding the unknown sides or angles of a triangle on the celestial sphere from the known sides and angles. One way of finding the time of day, for example, is to construct a triangle whose vertexes are the zenith, the north celestial pole and the sun's position. The observer must know the altitude of the sun and that of the pole; the former can be observed, and the latter is equal to the observer's latitude. The time is then given by the angle at the intersection of the meridian (the arc through the zenith and the pole) and the sun's hour circle (the arc through the sun and the pole).

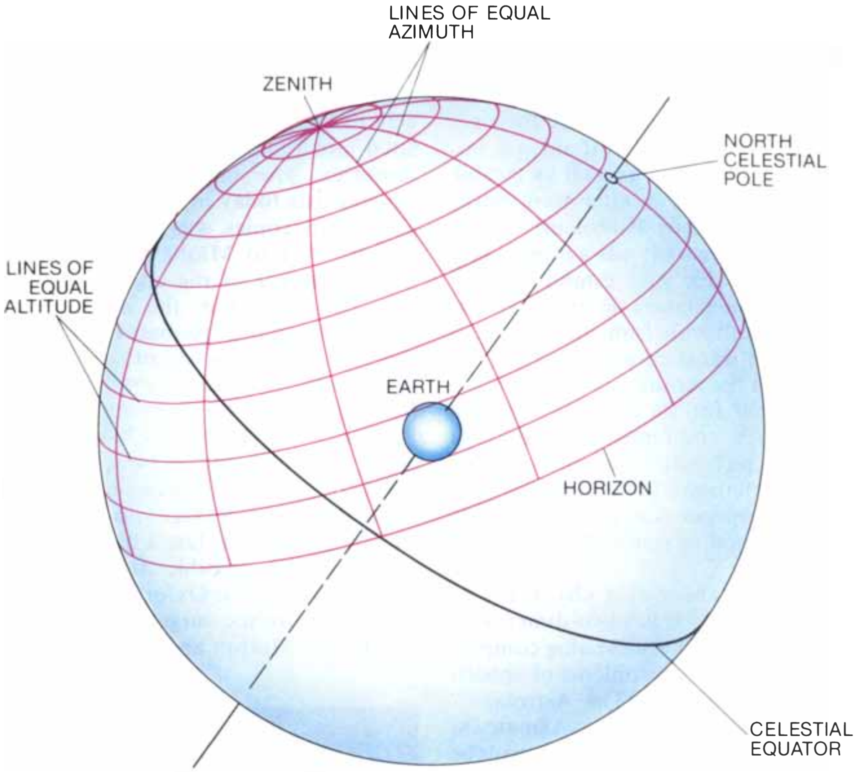
The method Ptolemy used to solve spherical triangles was a clumsy one devised late in the first century by Menelaus of Alexandria. It involved setting up two intersecting right triangles; by applying the Menelaus theorem it was possible to solve for one of the six sides, but only if the other five sides were known. To tell the time from the sun's altitude, for instance, repeated applications of the Menelaus theorem were required. For medieval Islamic astronomers there was an obvious challenge to find a simpler trigonometric method.

By the ninth century the six modern trigonometric functions—sine and cosine, tangent and cotangent, secant and cosecant—had been identified, whereas Ptolemy knew only a single chord function. Of the six, five seem to be essentially Arabic in origin; only the sine function was introduced into Islam from India. (The etymology of the word sine is an interesting tale. The Sanskrit word was *ardhajya,* meaning "half chord," which in Arabic was shortened and transliterated as *jyb.* In Arabic vowels are not spelled out, and so the word was read as *jayb,* meaning "pocket" or "gulf." In medieval Europe it was then translated as *sinus,* the Latin word for gulf.) From the ninth century onward the development of spherical trigonometry was rapid. Islamic astronomers discovered simple trigonometric identities, such as the law of sines, that made solving spherical triangles a much simpler and quicker process.

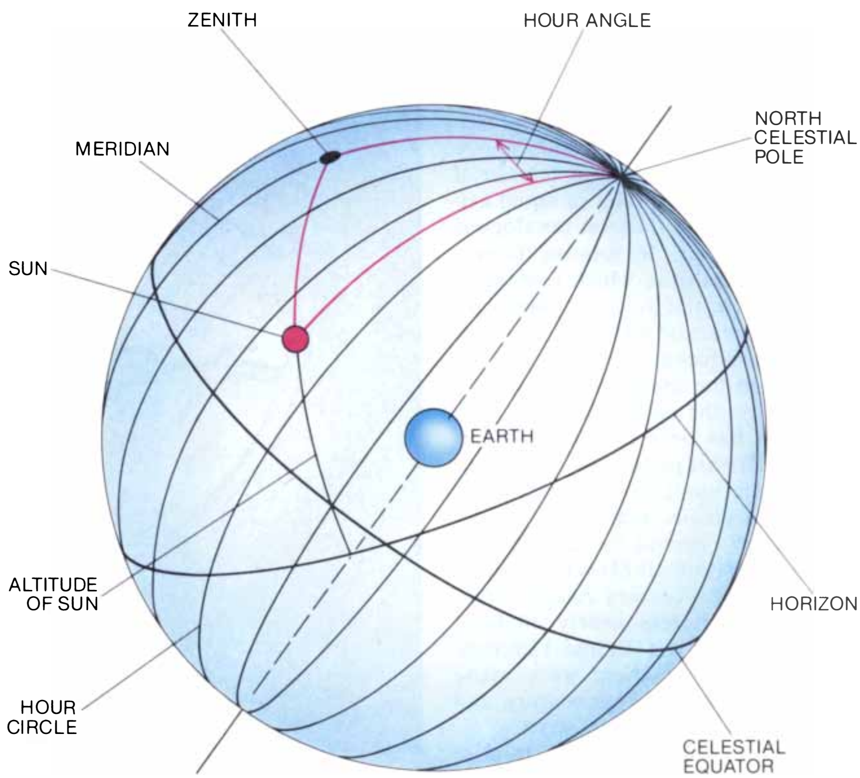
## Stars and Astrolabes

One of the most conspicuous examples of modern astronomy's Islamic heritage is the names of stars. Betelgeuse, Rigel, Vega, Aldebaran and Fomalhaut are among the names that are directly Arabic in origin or are Arabic translations of Ptolemy's Greek descriptions.

In the *Almagest* Ptolemy had provided a catalogue of more than 1,000 stars. The first critical revision of the catalogue was compiled by 'Abd al-Raḥmān al-Ṣūfī, a 10th-century Persian astronomer who worked in both




**ALTAZIMUTH COORDINATES map the celestial hemisphere visible at a given latitude. The sky is divided by lines of equal altitude above the horizon and lines of equal azimuth around the horizon. The latter converge toward the zenith, the point directly overhead.**




**SOLVING SPHERICAL TRIANGLES was a fundamental problem faced by Islamic astronomers. To tell time from the altitude of the sun, for example, they had to find the hour angle formed at the north celestial pole by the meridian (the great circle through the zenith and the poles) and the sun's hour circle (the great circle through the sun and the poles). At noon, when the sun crosses the meridian, its altitude is maximum and its hour angle is zero.**

Iran and Baghdad. Al-Ṣūfī's *Kitāb ṣuwar al-kawākib* ("Book on the Constellations of Fixed Stars") did not add or subtract stars from the *Almagest* list, nor did it remeasure their often faulty positions, but it did give improved magnitudes as well as Arabic identifications. The latter were mostly just translations of Ptolemy.

For many years it was assumed that al-Ṣūfī's Arabic had established the stellar nomenclature in the West. It now seems that his 14th- and 15th-century Latin translators went to a Latin version of the Arabic edition of Ptolemy himself for the star descriptions, which they combined with al-Ṣūfī's splendid pictorial representations of the constellations. Meanwhile the Arabic star nomenclature trickled into the West by another route: the making of astrolabes.

The astrolabe was a Greek invention. Essentially it is a two-dimensional model of the sky, an analog computer for solving the problems of spherical astronomy [see "The Astrolabe," by J. D. North; SCIENTIFIC AMERICAN, January, 1974]. A typical astrolabe consists of a series of brass plates nested in a brass matrix known in Arabic as the *umm* (meaning "womb"). The uppermost plate, called the *'ankabūt* (meaning "spider") or in Latin the rete, is an open network of two or three dozen pointers indicating the position of specific stars. Under the rete are one or more solid plates, each engraved with a celestial coordinate system appropriate for observations at a particular latitude: circles of equal altitude above the horizon (analogous to terrestrial latitude lines) and circles of equal azimuth around the horizon (analogous to longitude lines). By rotating the rete about a central pin, which represents the north celestial pole, the daily motions of the stars on the celestial sphere can be reproduced.

Although the astrolabe was known in antiquity, the earliest dated instrument that has been preserved comes from the Islamic period [*see cover of this issue*]. It was made by one Nastulus in 315 of the Hegira era (A.D. 927–28), and it is now one of the treasures of the Kuwait National Museum. Only a handful of 10th-century Arabic astrolabes exist, whereas nearly 40 have survived from the 11th and 12th centuries. Several of these were made in Spain in the mid-11th century and have a distinctly Moorish style.

The earliest extant Arabic treatise on the astrolabe was written in Baghdad by one of Caliph al-Ma'mūn's astronomers, 'Alī ibn 'Īsā. Later members of the Baghdad school, notably al-Farghānī, also wrote on the astrolabe. Al-Farghānī's treatise was impressive for the mathematical way he applied the instrument to problems in astrology, astronomy and timekeeping.

Many of these treatises found their way to Spain, where they were translated into Latin in the 12th and 13th centuries. The most popular work, which exists today in about 200 Latin manuscript copies, was long mistakenly attributed to Māshā'allāh, a Jewish astronomer of the eighth century who participated in the decision to found Baghdad; it probably is a later pastiche from a variety of sources. In about 1390 this treatise was the basis for an essay on the astrolabe by the English poet Geoffrey Chaucer. Indeed, England seems to have been the gateway for the introduction of the astrolabe from Spain into Western Christendom in the late 13th and 14th centuries. It is possible that scientific activity centered at Oxford at the time contributed to the surge of interest in the device. Merton and Oriel colleges of the University of Oxford still own fine 14th-century astrolabes.

On them one finds typical sets of Arabic star names written in Gothic Latin letters. Included on the Merton College astrolabe, for example, are Arabic names that have evolved into standard modern nomenclature: Wega, Altahir, Algeuze, Rigil, Elfeta, Alferaz and Mirac. Thus as a result of the astrolabe tradition of Eastern Islam, transmitted through Spain to England, most navigational stars today have Arabic names, either indigenous ones or Arabic translations of Ptolemy's Greek descriptions.

## Refining Ptolemy

It would be wrong to conclude from the preponderance of Arabic star names that Islamic astronomers made exhaustive studies of the sky. On the contrary, their observations were quite limited. For instance, the spectacular

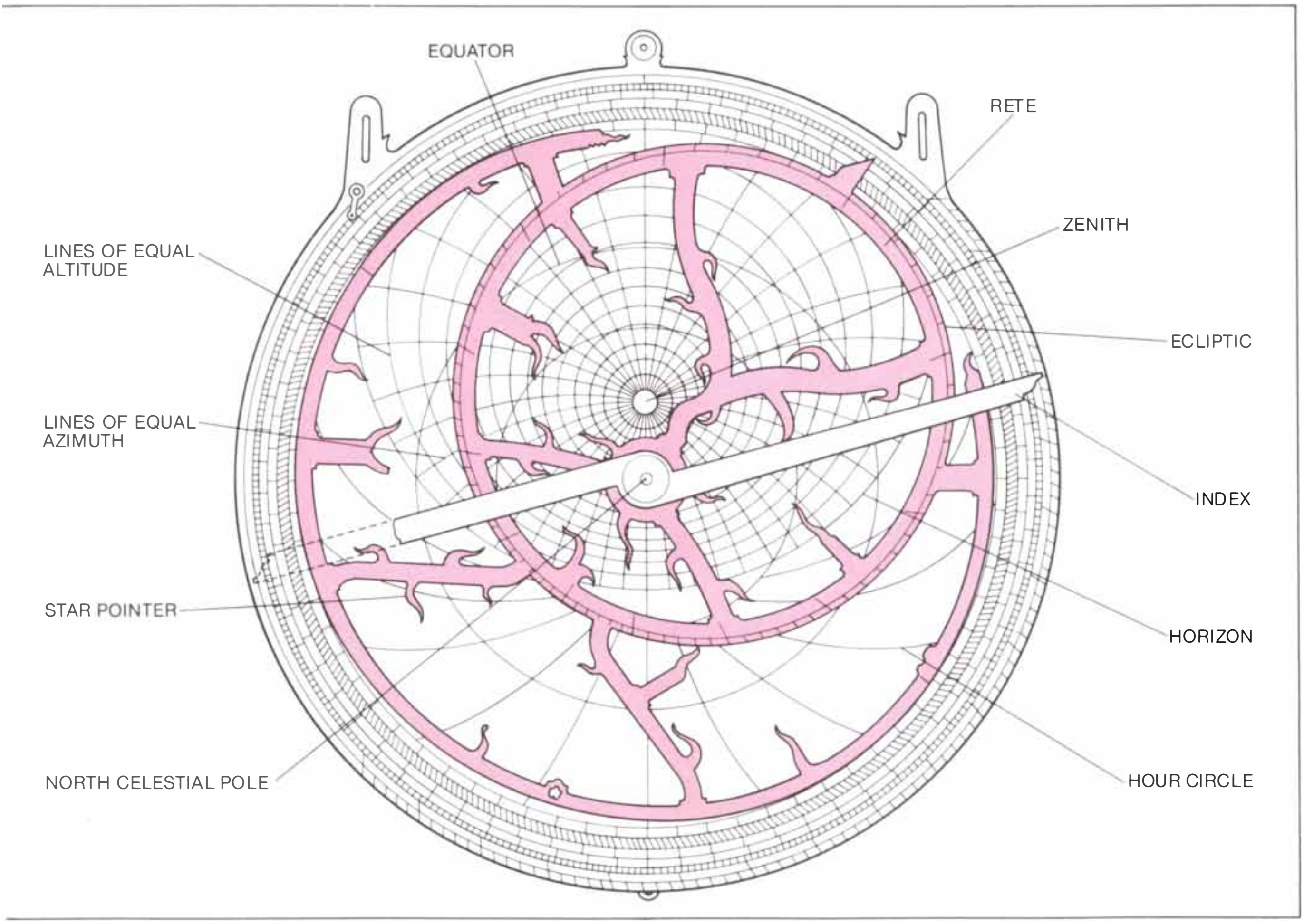
**ASTROLABE simplified astronomical calculations, including the telling of time. Its nested brass plates are a projection of the celestial sphere onto two dimensions. The top plate, called the rete, is an open network of pointers indicating the positions of prominent stars. The off-center circle on the rete is the ecliptic. Under the rete is a solid plate on which is engraved a celestial coordinate system for a particular latitude: lines of equal altitude, lines of equal azimuth and hour circles. By turning the rete around the central pin, which represents the**

supernova (stellar explosion) of 1054, which produced the Crab Nebula, went virtually unrecorded in Islamic texts even though it was widely noted in China. Modern astronomers struck by this glaring gap often do not realize that Islamic astronomers failed to document most specific astronomical phenomena. They had little incentive to do so. Their astrology, unlike that of the Chinese, depended not so much on unusual heavenly omens as on planetary positions, and these were quite well described by the Ptolemaic procedures.

The planetary models that Ptolemy devised in the second century A.D. had the sun, the moon and the planets moving around the earth. A simple circular orbit, however, could not account for the fact that a planet periodically seems to reverse its direction of motion across the sky. (According to the modern heliocentric viewpoint, this apparent retrograde motion occurs when the earth is passing or being passed by another planet on its way around the sun.) Hence Ptolemy had each planet moving on an epicycle, a rotating circle whose center moved about the earth on a larger circle called the deferent. The epicycle, together with other geometric devices invented by Ptolemy, gave a fairly good first approximation to the apparent motion of the planets. As a great theoretician, Ptolemy must have been fairly confident of the particular geometry of his models, since he never described how he settled on it.

On the other hand, the idea of applying mathematics to a specific numerical description of the physical world was something rather novel for the Hellenistic Greeks, quite different from the pure mathematics of Euclid and Apollonius. In this part of his program Ptolemy must have realized that improved values for the numerical parameters of his models were both desirable and inevitable, and so he gave careful instructions on how to establish the parameters from a limited number of selected observations. The Islamic astronomers learned this lesson all too well. They limited their observations, or at least the few they chose to record, primarily to measurements that could be used for rederiving key parameters. These included the orientation and eccentricity of the solar orbit and the inclination of the ecliptic plane.

An impressive example of an Islamic astronomer working strictly within a Ptolemaic framework but establishing new values for Ptolemy's parameters was Muḥammad al-Bāttanī, a younger contemporary of Thābit ibn Qurra. Al-Bāttanī's *Zīj* ("Astronomical Tables") is still admired as one of the most important astronomical works between the time of Ptolemy and that of Copernicus. Among other things, al-Bāttanī was able to establish the position of the solar orbit (equivalent in

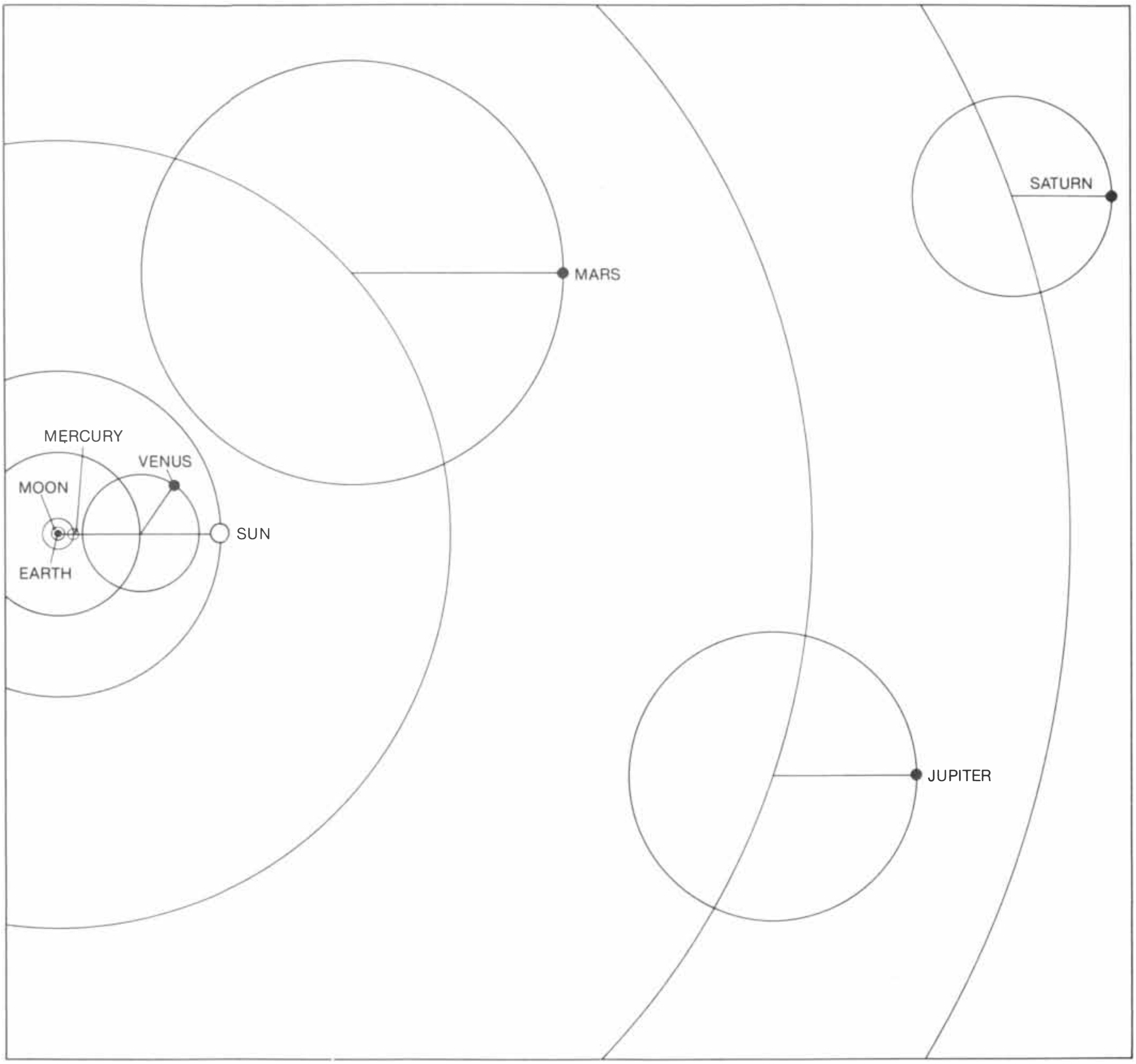



**north celestial pole, one could reproduce the daily motions of the stars. The first step in telling the time was to determine the altitude of the sun (or of a star) with the help of a sighting bar and a degree scale on the back of the astrolabe. Next the rete was turned until the position of the sun on the ecliptic (or the star pointer) was at the correct altitude line. The hour angle was then given by the angle at the central pin between the meridian line and a line through the object's position. The astrolabe shown here was made in England in the 14th century and belongs to Merton College of the University of Oxford. Astrolabes were the chief means by which Arabic star names were transmitted to the West. Names such as Wega, Altahir, Algeuze, Rigil, Elfeta and Alferaz are visible in the photograph.**

modern terms to finding the position of the earth's orbit) with better success than Ptolemy had achieved.

Because al-Bāttanī does not describe his observations in detail, it is not clear whether he adopted an observational strategy different from that of Ptolemy. In any case his results were good, and centuries later his parameters for the solar orbit were widely known in Europe. His *Zīj* first made its way to Spain. There it was translated into Latin early in the 12th century and into Castilian a little more than 100 years later. The fact that only a single Arabic manuscript copy survives (in the Escorial Library near Madrid) suggests that al-Bāttanī's astronomy was not as highly regarded in Islam as it was in Europe, where the advent of printing ensured its survival and in particular made it available to Copernicus and his contemporaries. In *De revolutionibus orbium coelestium* ("On the Revolutions of the Heavenly Spheres") the Polish astronomer mentions his ninth-century Muslim predecessor no fewer than 23 times.

In contrast, one of the greatest astronomers of medieval Islam, 'Alī ibn 'Abd al-Raḥmān ibn Yūnus, remained completely unknown to European astronomers of the Renaissance. Working in Cairo a century after al-Bāttanī, Ibn Yūnus wrote a major astronomical handbook called the *Ḥākimī Zīj*. Unlike other Arabic astronomers, he prefaced his *Zīj* with a series of more than 100 observations, mostly of eclipses and planetary conjunctions. Although Ibn Yūnus' handbook was widely used in Islam, and his timekeeping tables




**PTOLEMAIC SYSTEM placed the sun, the moon and the planets in orbit around the earth. The moon and the five known planets moved on subsidiary circles, called epicycles, whose centers rode on larger circles called deferents. The epicycles reproduced the apparent retrograde motion of the planets (*right*). Even outside the retrograde loops, however, the orbital speed of a planet varies. To represent this nonuniform motion Ptolemy adopted two further geometric devices. First, he made the deferents eccentric, that is, he moved their centers away from the earth. Second, he assumed that the angular motion of a planet was uniform around the equant, a**

survived in use in Cairo into the 19th century, his work became known in the West less than 200 years ago.

Throughout the entire Islamic period astronomers stayed securely within the geocentric framework. For this one should not criticize them too harshly. Until Galileo's telescopic observations of the phases of Venus in 1610, no observational evidence could be brought against the Ptolemaic system. Even Galileo's observations could not distinguish between the geo-heliocentric system of Tycho Brahe (in which the other planets revolved about the sun but the sun revolved about the earth) and the purely heliocentric system of Copernicus [see "The Galileo Affair," by Owen Gingerich; SCIENTIFIC AMERICAN, August, 1982]. Furthermore, although Islamic astronomers followed Ptolemy's injunction to test his results, they did not limit themselves simply to improving his parameters. The technical details of his models were not immune from criticism. These attacks, however, were invariably launched on philosophical rather than on observational grounds.

## Doubting Ptolemy

Ptolemy's models were essentially a mathematical system for predicting the positions of the planets. Yet in the *Planetary Hypotheses* he did try to fit the models into a cosmological system, the Aristotelian scheme of tightly nested spheres centered on the earth. He placed the nearest point of Mercury's path immediately beyond the most distant point of the moon's path; immediately beyond the farthest excursion of Mercury lay the nearest approach of Venus, and so on through the spheres for the sun, Mars, Jupiter and Saturn.

To reproduce the observed nonuniform motions of the planets, however, Ptolemy adopted two purely geometric devices in addition to the epicycle. First, he placed the deferent circles off-center with respect to the earth. Second, he made the ingenious assumption that the motion of celestial bodies was uniform not around the earth, nor around the centers of their deferents, but instead around a point called the equant that was opposite the earth from the deferent center and at an equal distance. Eccentric deferents and equants did a good job of representing the varying speeds with which planets are seen to move across the sky, but to some minds they were philosophically offensive.

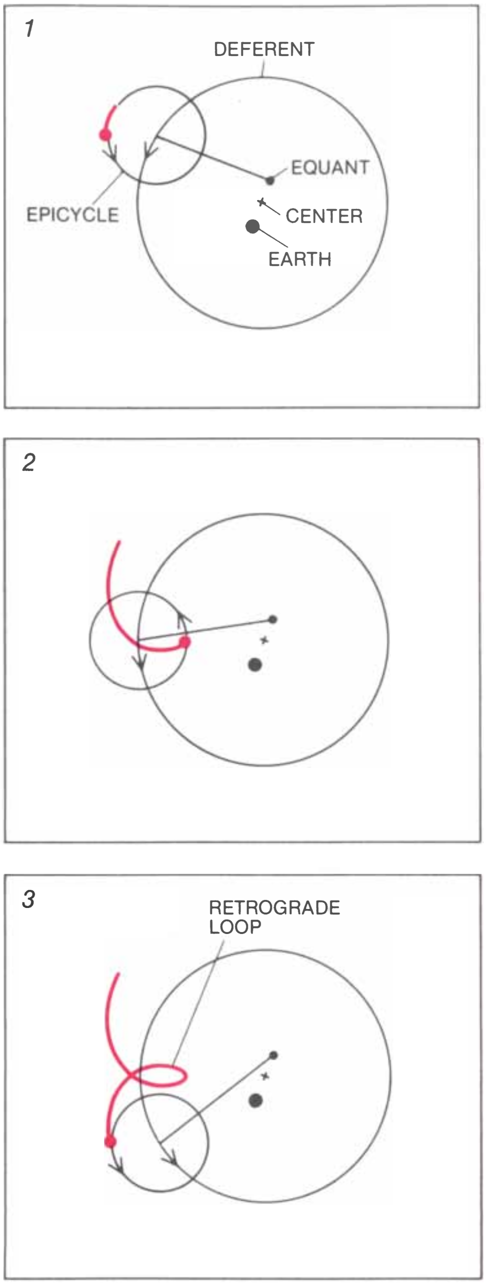
The equant in particular was objectionable to philosophers who thought of planetary spheres as real physical objects, each sphere driven by the one outside it (and the outermost driven by the prime mover), and who wanted to be able to construct a mechanical model of the system. For example, as was pointed out by Maimonides, a Jewish scholar of the 12th century who worked in Spain and Cairo, the equant point for Saturn fell right on the spheres for Mercury. This was clearly awkward from a mechanical point of view. Furthermore, the equant violated the philosophical notion that heavenly bodies should be moved by a system of perfect circles, each of which rotated with uniform angular velocity about its center. To some purists even Ptolemy's eccentric deferents, which moved the earth away from the center of things, were philosophically unsatisfactory.

The Islamic astronomers adopted the Ptolemaic-Aristotelian cosmology, but eventually criticism emerged. One of the first critics was Ibn al-Haytham (Alhazen), a leading physicist of 11th-century Cairo. In his *Doubts on Ptolemy* he complained that the equant failed to satisfy the requirement of uniform circular motion, and he went so far as to declare the planetary models of the *Almagest* false.

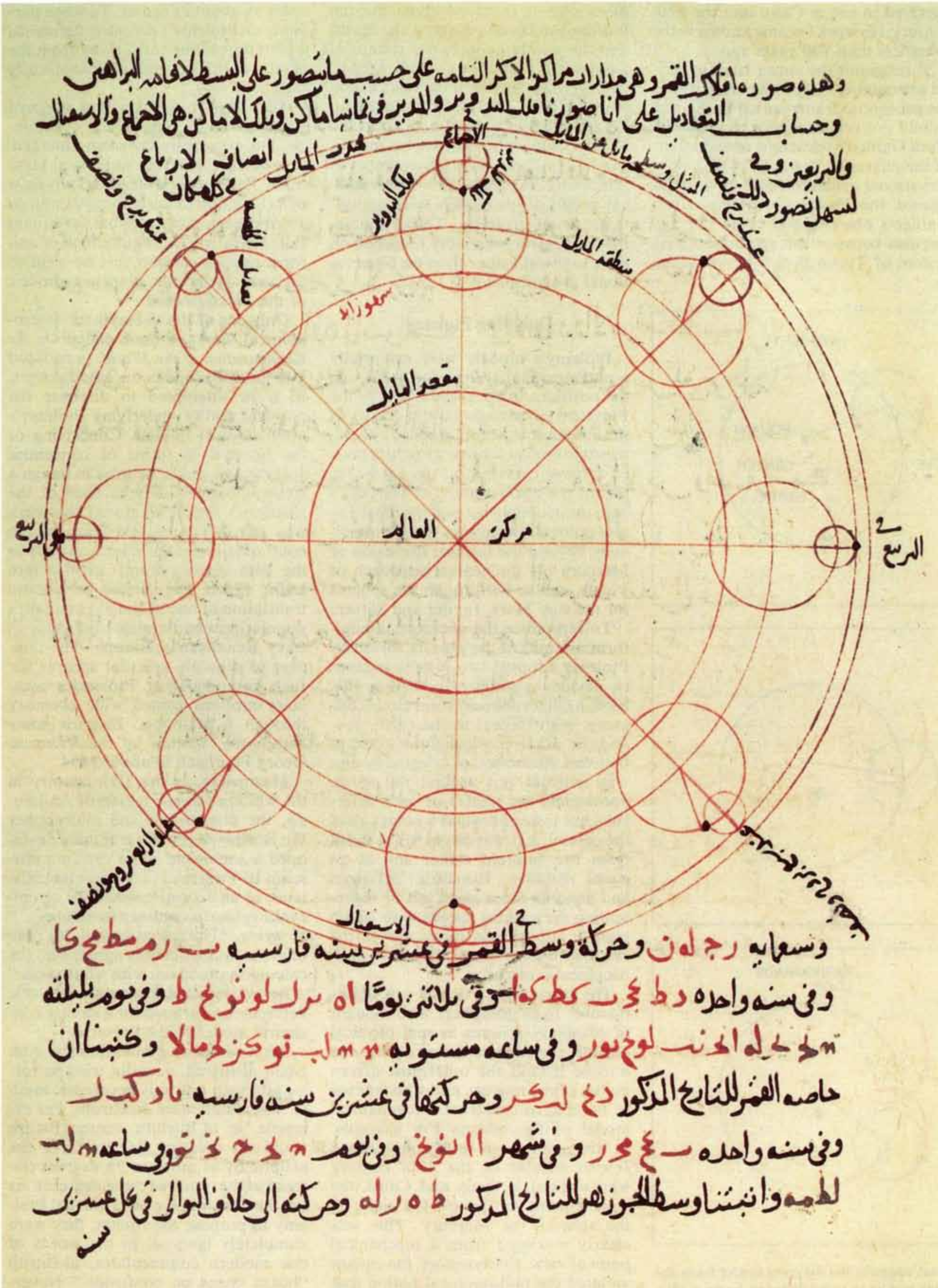
Only one of Ibn al-Haytham's astronomical works, a book called *On the Configuration of the World,* penetrated into Latin Europe in the Middle Ages. In it he attempted to discover the physical reality underlying Ptolemy's mathematical models. Conceiving of the heavens in terms of concentric spheres and shells, he tried to assign a single spherical body to each of the *Almagest's* simple motions. The work was translated into Castilian in the court of Alfonso the Wise, and early in the 14th century from Castilian into Latin. Either this version or a Latin translation of one of Ibn al-Haytham's popularizers had a major influence in early Renaissance Europe. The concept of separate celestial spheres for each component of Ptolemy's planetary motions gained wide currency through a textbook, *Theorica novae planetarum,* written by the Viennese Georg Peurbach in about 1454.

Meanwhile, in the 12th century in the western Islamic region of Andalusia, the astronomer and philosopher Ibn Rushd (Averroës) gradually developed a somewhat more extreme criticism of Ptolemy. "To assert the existence of an eccentric sphere or an epicyclic sphere is contrary to nature . . . ," he wrote. "The astronomy of our time offers no truth, but only agrees with the calculations and not with what exists." Averroës rejected Ptolemy's eccentric deferents and argued for a strictly concentric model of the universe.

An Andalusian contemporary, Abu Isḥāq al-Biṭrūjī, actually tried to formulate such a strictly geocentric model. The results were disastrous. For example, in al-Biṭrūjī's system Saturn could on occasion deviate from the ecliptic by as much as 26 degrees (instead of the required three degrees). As for the observed motions that led Ptolemy to propose the equant, they were completely ignored. In the words of one modern commentator, al-Biṭrūjī "heaps chaos on confusion." Nevertheless, early in the 13th century his work was translated into Latin under the name Alpetragius, and from about




**point opposite the deferent center from the earth and at an equal distance. Some medieval Islamic astronomers considered Ptolemy's equants as well as his eccentric deferents to be philosophically objectionable.**




**DOUBLE-EPICYCLE SYSTEM proposed in the 14th century by Ibn al-Shāṭir eliminated Ptolemy's equants and put the sun, moon and planets in concentric orbits around the earth. This diagram, from a manuscript now at Oxford, shows positions of the moon.**

1230 on his ideas were widely discussed throughout Europe. Even Copernicus cited his order of the planets, which placed Venus beyond the sun.

At the other end of the Islamic world a fresh critique of the Ptolemaic mechanisms was undertaken in the 13th century by Naṣīr al-Dīn al-Ṭūsī. One of the most prolific Islamic polymaths, with 150 known treatises and letters to his credit, al-Ṭūsī also constructed a major observatory at Maragha (the present-day Marāgheh in Iran).

Al-Ṭūsī found the equant particularly dissatisfactory. In his *Tadhkira* ("Memorandum") he replaced it by adding two more small epicycles to the model of each planet's orbit. Through this ingenious device al-Ṭūsī was able to achieve his goal of generating the nonuniform motions of the planets by combinations of uniformly rotating circles. The centers of the deferents, however, were still displaced from the earth. Two other astronomers at the Maragha observatory, Mu'ayyad al-Dīn al-'Urḍī and Quṭb al-Dīn al-Shīrāzī, offered an alternative arrangement, but this system too retained the philosophically objectionable eccentricity.

Finally a completely concentric rearrangement of the planetary mechanisms was achieved by Ibn al-Shāṭir, who worked in Damascus in about 1350. By using a scheme related to that of al-Ṭūsī, Ibn al-Shāṭir succeeded in eliminating not only the equant but also certain other objectionable circles from Ptolemy's constructions. He thereby cleared the way for a perfectly nested and mechanically acceptable set of celestial spheres. (He described his work thus: "I found that the most distinguished of the later astronomers had adduced indisputable doubts concerning the well-known astronomy of the spheres according to Ptolemy. I therefore asked Almighty God to give me inspiration and help me to invent models that would achieve what was required, and God—may He be praised and exalted—did enable me to devise universal models for the planetary motions in longitude and latitude and all other observable features of their motions, models that were free from the doubts surrounding previous ones.") Yet Ibn al-Shāṭir's solution, along with the work of the Maragha astronomers, remained generally unknown in medieval Europe.

## Influence on Copernicus?

Ibn al-Shāṭir's forgotten model was rediscovered in the late 1950's by E. S. Kennedy and his students at the American University of Beirut. The discovery raised an intriguing question. It was quickly recognized that the Ibn al-Shāṭir and Maragha inventions were the same type of mechanism used by Copernicus a few centuries later to eliminate the equant and to generate the intricate changes in the position of the earth's orbit. Copernicus, of course, adopted a heliocentric arrangement, but the problem of accounting for the slow but regular changes in a planet's orbital speed remained exactly the same. Since Copernicus agreed with the philosophical objections to the equant—like some of his Islamic predecessors, he apparently believed celestial motions were driven by physical, crystalline spheres—he too sought to replace Ptolemy's device. In a preliminary work, the *Commentariolus,* he employed an arrangement equivalent to Ibn al-Shāṭir's. Later, in *De revolutionibus,* he reverted to the use of eccentric orbits, adopting a model that was the sun-centered equivalent of the one developed at Maragha.

Could Copernicus have been influenced by the Maragha astronomers or by Ibn al-Shāṭir? No Latin translation has been found of any of their works or indeed of any work describing their models. It is conceivable that Copernicus saw an Arabic manuscript while he was studying in Italy (from 1496 to 1503) and had it translated, but this seems highly improbable. A Greek translation of some of the al-Ṭūsī material is known to have reached Rome in the 15th century (many Greek manuscripts were carried west after the fall of Constantinople in 1453), but there is no evidence that Copernicus ever saw it.

Scholars are currently divided over whether Copernicus got his method for replacing the equant by some unknown route from the Islamic world or whether he found it on his own. I personally believe he could have invented the method independently.

Nevertheless, the whole idea of criticizing Ptolemy and eliminating the equant is part of the climate of opinion inherited by the Latin West from Islam. The Islamic astronomers would probably have been astonished and even horrified by the revolution started by Copernicus. Yet his motives were not completely different from theirs. In eliminating the equant, and even in placing the planets in orbit around the sun, Copernicus was in part trying to formulate a mechanically functional system, one that offered not only a mathematical representation but also a physical explanation of planetary motions. In a profound sense he was simply working out the implications of an astronomy founded by Ptolemy but transformed by the Islamic astronomers. Today that heritage belongs to the entire world of science.

# Internal Rhythms in Bird Migration

*Migratory birds have a clock that tells them when to begin and end their flight. It is based on rhythms with a period of about a year. Remarkably, the clock also helps the birds to find their destinations*

by Eberhard Gwinner

The sight of a flock of birds migrating south in the fall or north in the spring hardly ever fails to evoke a sense of wonder. The flight may be the orderly aerobatics of a V of Canada geese or the ragtag progress of a group of starlings. Whatever its details, the overwhelming impression conveyed to the observer is that of a powerful inner impulse. The birds do not hesitate in their flight; they travel smoothly and unerringly toward a goal far out of the viewer's sight. Where does the impulse come from that guides the birds toward warmer climates in winter and brings them back to their northern breeding grounds in the spring?

There are two possible kinds of answer to the question. The first is that the impulse springs from factors originating outside the bird, in its environment. Among those that come to mind immediately are the changes in temperature and in the duration of daylight that accompany the transition from season to season. A decrease in day length or in temperature might be sufficient to trigger the physiological responses that send the bird south; increases in those factors might trigger the converse physiological changes, sending the bird north again. The second general kind of explanation is that the impulse comes from within the bird.

In the past most research on the control of bird migration has concentrated on the influence of external factors. For the past 20 years, however, my colleagues and I have been studying the effect of internal rhythms on the migratory pattern. We have concluded that endogenous rhythms with a period of about a year have a powerful influence on when and how migration takes place. Such "circannual" cycles seem to provide an overall framework for the timing of migration. In addition they affect the details of the migratory flight, perhaps even helping the birds to navigate toward specific targets. Circannual rhythms can be modified by environmental factors and by learning. Nevertheless, in identifying them it seems we have discovered an essential component of the mechanism that controls avian migration.

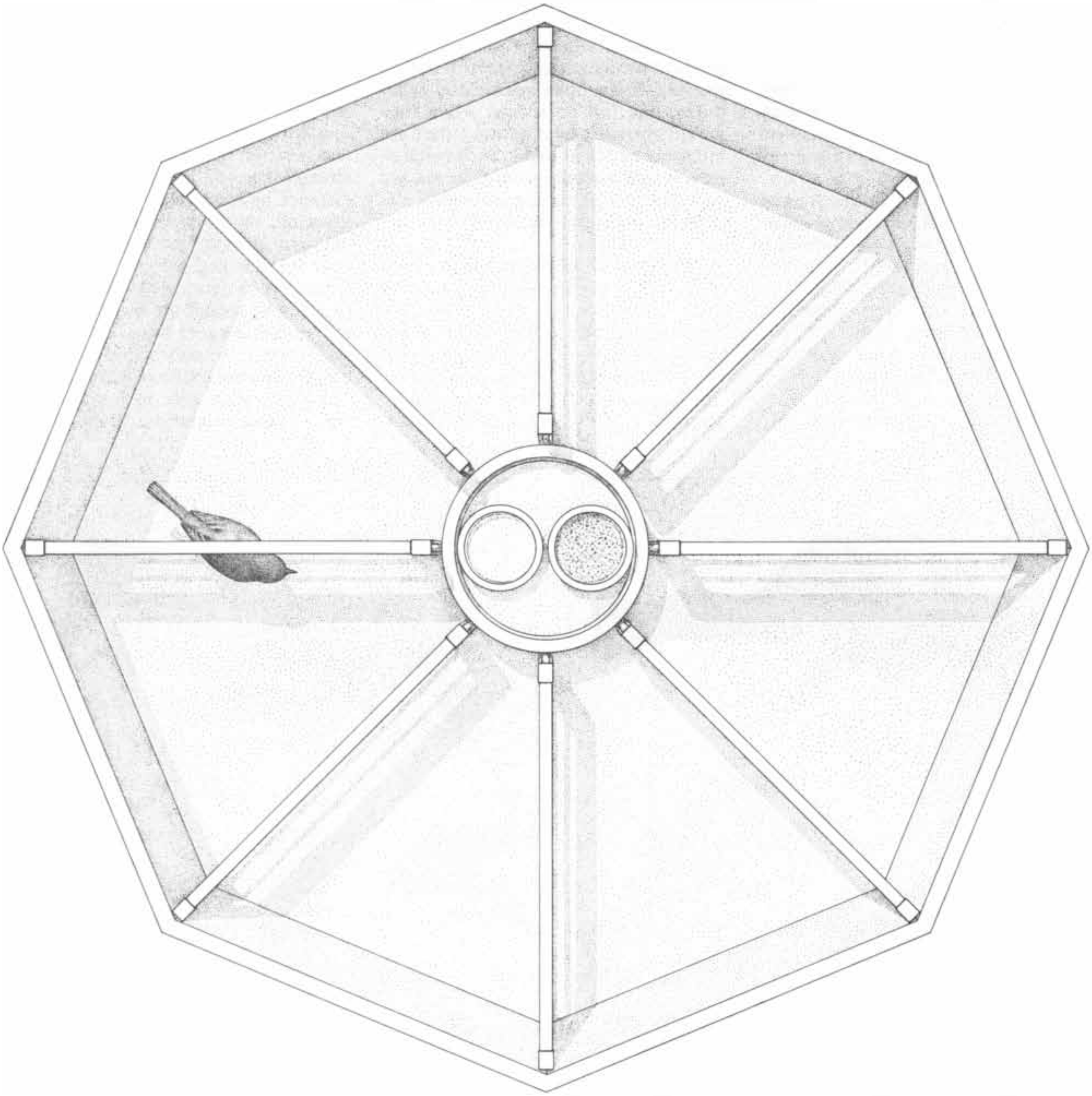
Since our institute, the Vogelwarte Radolfzell, lies in southern Germany, it was natural for us to study species that participate in the Palaearctic-African system of bird migrations. The Palaearctic-African system is one of several great systems of avian migration, each of which encompasses the flight paths of many species. A second migratory system connects North America with Central and South America; a third stretches from northeastern Asia to southeastern Asia and Australia. Since a large proportion of the birds that live in temperate areas migrate every year to southern regions, these migratory systems can be very large and complex. Of them, the Palaearctic-African one has been the most closely studied.

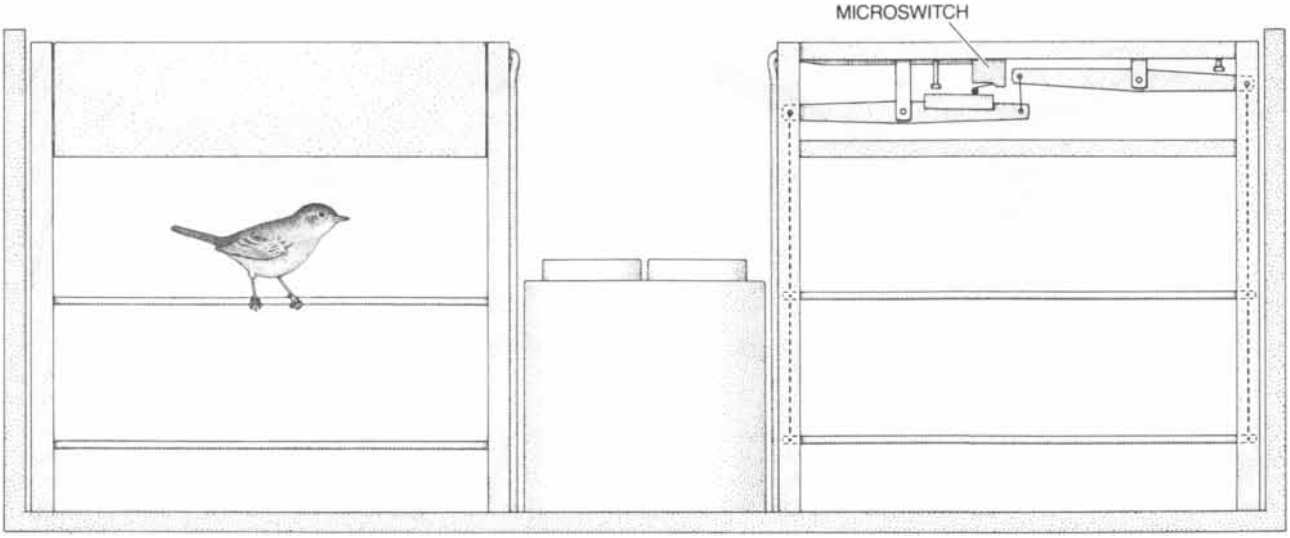
Much of what is known about the Palaearctic-African system has come from bird-banding experiments that began in Europe around the turn of the century. More than 60 million birds have been banded in Europe over the years, and the recovery of more than one million of them has yielded considerable information about their routes and wintering grounds. Partly on the basis of such information R. E. Moreau of the Edward Grey Institute in England has estimated that every year more than five billion birds invade Africa from Europe. The wintering grounds of the invaders extend from northern Africa to the Cape of Good Hope, and the distances they cover vary greatly. Whereas a woodcock may travel only a few hundred kilometers across the Mediterranean, a Siberian ruff may have to fly more than 12,000 kilometers across Asia and eastern Europe before reaching its winter home in central Africa.

For the most part our experiments have been carried out with European warblers (of the family Sylviidae) and flycatchers (of the family Muscicapidae). In studying how migration is controlled among birds of these species there were three main questions that had to be answered. The first was how the timing of migration is controlled. The second was how the bird is able to navigate to a specific target area on each leg of the migration. The third was how the organism can withstand the great energetic demands imposed by long-distance flying, particularly across oceans or deserts, where the bird cannot easily replenish depleted energy reserves.

All three questions are interesting, but investigating them simultaneously would have made for a somewhat unwieldy project. Accordingly our research strategy did not give the three questions equal attention. Instead, from the beginning of our work in the mid-1960's we concentrated on migrational timing. Specifically, we wanted

**ORIENTATION CAGE was used by the author to test the directional preferences of migratory birds at various times of the year. The cage has eight pairs of perches (*upper panel*). Each perch is connected by a string to a microswitch (*lower panel*). When the bird jumps on one of the perches, the microswitch completes a circuit and a nearby event recorder (not shown) registers the electrical impulses. The apparatus makes it possible to find out which part of the cage (and thus which compass direction) the bird prefers. By this means the author was able to show that the directional preferences of migratory birds undergo seasonal changes as the result of endogenous rhythms. The bird studied in the author's experiments was the garden warbler (*Sylvia borin*), and that is the species shown in the apparatus.**



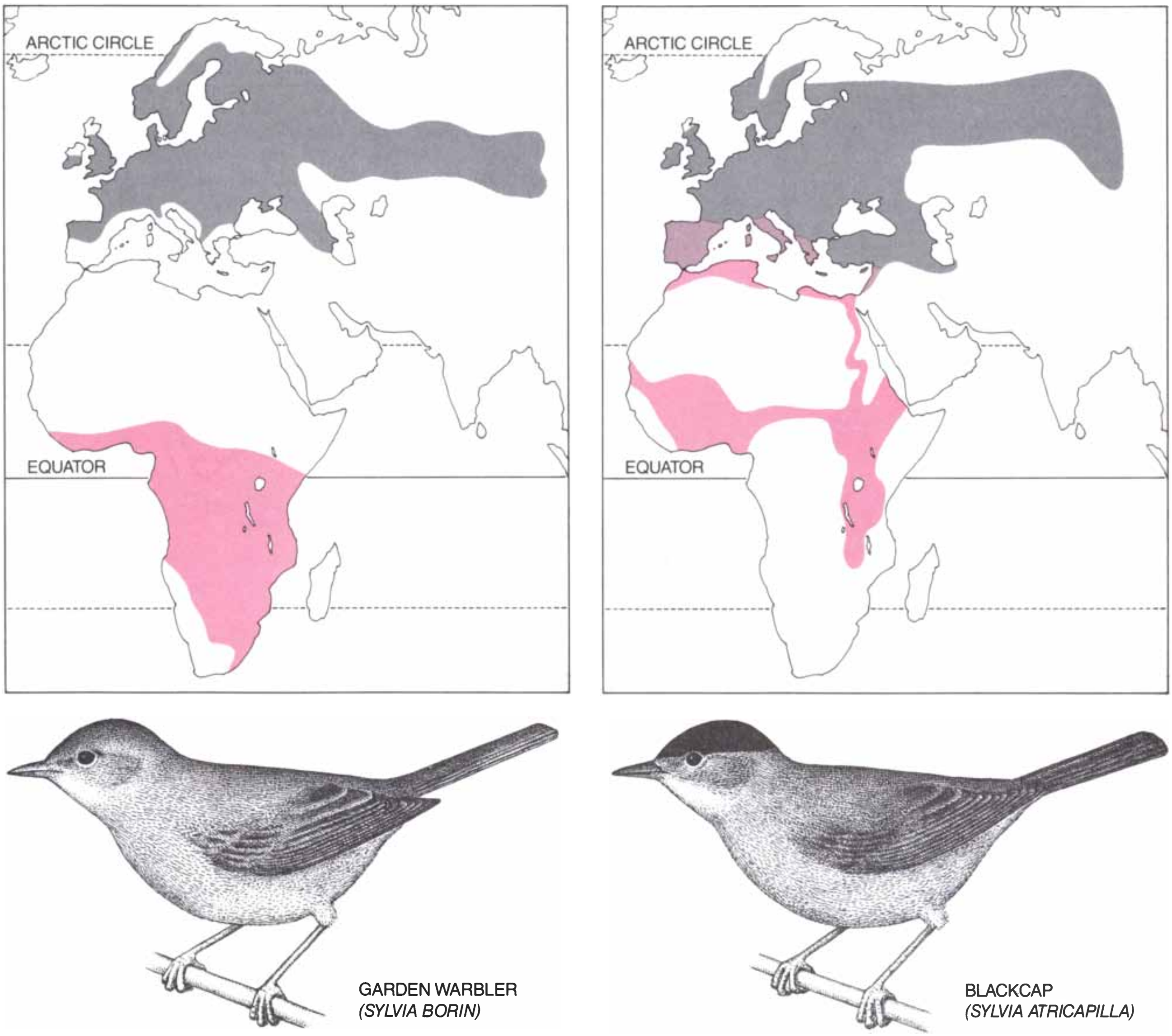

MICROSWITCH

to identify the source of the information that causes the bird to prepare for migration, begin its travels and ultimately terminate the migratory flight. In attempting to understand the problem of timing, however, we also obtained results that can aid in answering the other two questions.

Ultimately we carried out a long series of experiments aimed at elucidating the control of migratory timing. In many of them caged birds were carefully observed in order to detect the appearance of behaviors associated with migration. Such behaviors were of great significance in our work, because they served as indicators of the successive stages of the annual migratory cycle. One example is "migratory restlessness." The warblers and flycatchers we studied are nocturnal migrants: they fly at night and rest during the day. If the birds are held in cages during the fall or spring, when they would normally be migrating, they are intensely active at night. Such restlessness, which can take several forms, is a sign that the physiological events underlying migration have begun.

Migratory restlessness had a crucial role in the first experiment of the series, conducted in 1966. The goal of that work was to find out how the spring migration is initiated in the willow warbler (*Phylloscopus trochilus*). The willow warbler is a small, inconspicuous bird. Its breeding ground extends over much of central and northern Eurasia, and its wintering grounds are in tropical regions of the Southern Hemisphere. The willow warblers we worked with breed in Germany and spend the winter in equatorial Africa.

When we began our work, it was already known that for some short-distance migrants changes in the photoperiod, or the average duration of daylight, provide an important stimulus for the beginning of migration. Of course in equatorial Africa, where the warbler spends the winter, the photoperiod undergoes little seasonal variation. Since the photoperiod varies so little, it seemed unlikely that light cues had much of a role in initiating the bird's spring migration. It was tempt-




**TWO SPECIES of warblers are among the birds studied by the author. At the left is the garden warbler and at the right is the blackcap (*Sylvia atricapilla*). Both are small and visually inconspicuous, but they are excellent songsters. Above each bird is a map depicting its breeding range (*gray*) and winter range (*color*). In fall the birds leave their breeding ranges and head south. Garden warblers spend the winter in central and southern Africa. Most blackcaps winter in southern Europe or in Africa north of the Equator; some blackcaps from eastern populations migrate to southern Africa. In spring the birds return to Europe and northern Asia to breed.**

ing to hypothesize that an internal timing mechanism triggers the homeward journey with no assistance from external stimuli.

To test that hypothesis I investigated the behavior of three groups of willow warblers. All three groups were held in experimental cages that made it possible to measure the degree of their migratory restlessness. In most warblers the nocturnal activity generally takes the form of hopping: the birds leap on and off their perches and jump around the cage almost continuously. Microswitches were mounted under one of the perches in the experimental cage. When the bird hopped on the perch, the microswitch completed a circuit and caused a nearby event recorder to make a mark on a strip of paper. By counting the number of intervals (half hours, say) in which the bird was active, it was possible to measure the intensity of the restlessness.

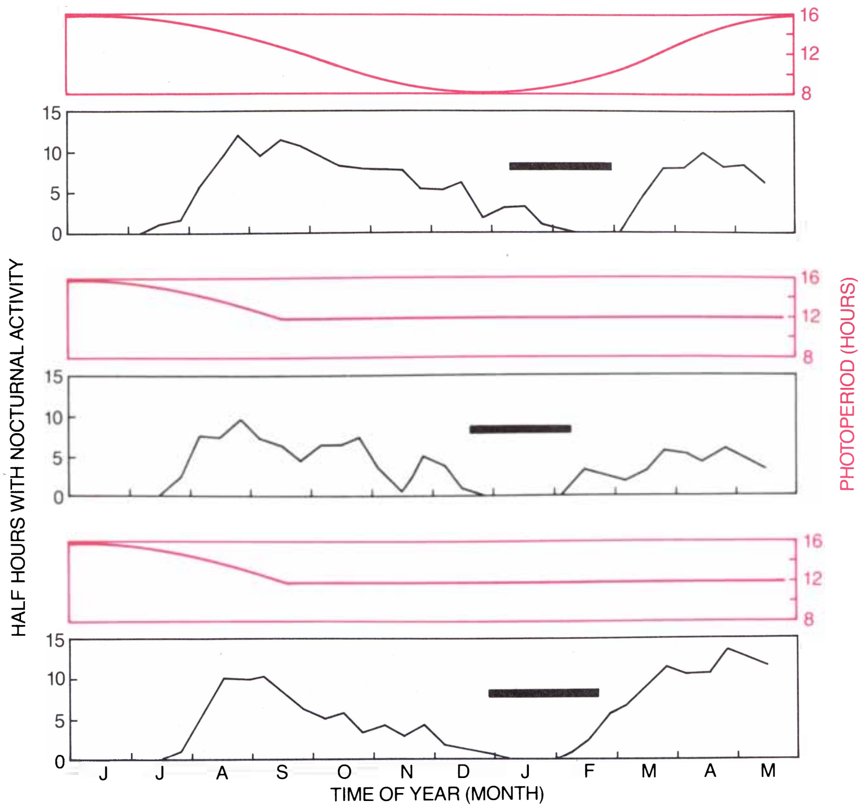
One group of birds was held in our laboratory at a constant temperature and under a constant photoperiod of 12 hours per day. The aim of the experiment was to see whether those birds would undergo the seasonal events associated with migration on the normal schedule without external cues. As a control, a second group was transported to a place in eastern Zaire at a latitude of about two degrees south that lies within the bird's normal wintering range. A third group was held throughout the winter at our institute in a room with large windows and was thereby exposed to the variation in photoperiod typical of temperate regions. The three groups were held in individual cages to determine the end of the fall migratory restlessness and its onset in the spring.

The results of the experiment were quite clear-cut. In all three groups the fall migratory restlessness ended during December and January and the spring restlessness began during February and March. That is the pattern typical of willow warblers in the wild. Furthermore, each group molted (renewed its plumage) at the correct time. Molting patterns vary widely among bird species. For each species, however, the molt takes place at a fixed time in relation to the migratory flight. Free-living willow warblers carry out a full molt in midwinter, just after reaching their wintering grounds, and that is when all three groups of caged birds molted.

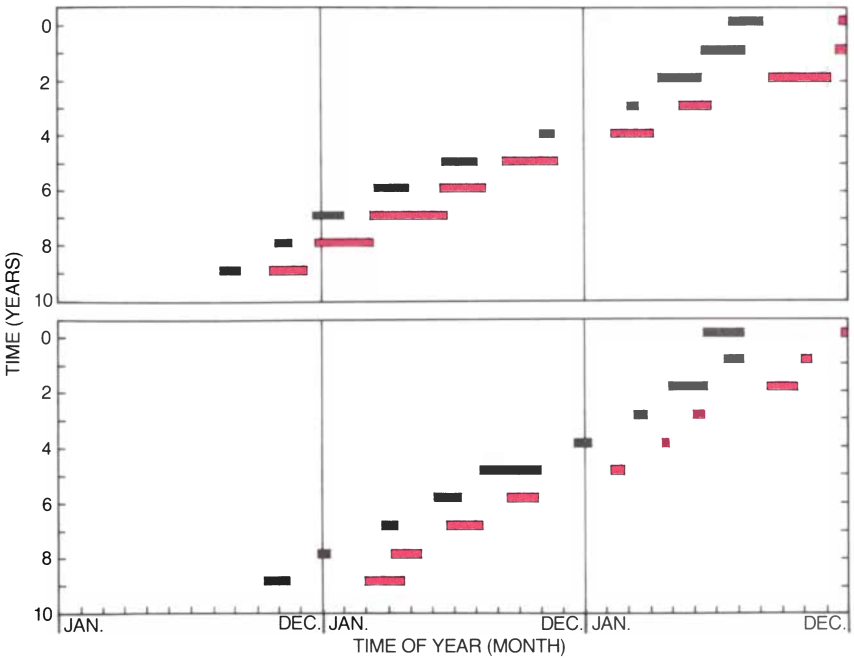
The fact that warblers under constant conditions of light and temperature adhere to the normal schedule of molting and restlessness suggested that the birds exploit internal cues to time their seasonal activities. The precise timing mechanism was unclear, however, because the results of the willow warbler experiment could be interpreted two ways. The correct timing of the events in winter might be due to an internal mechanism that is set in motion each spring in the breeding area. Such a mechanism would resemble an hourglass in that it would run for only a year before needing to be reset by external stimuli. On the other hand, the scheduling of seasonal events might reflect the operation of a mechanism that is more like a clock: an endogenous annual rhythm that continues to operate for many years without needing to be restarted.

The 1966 experiment did not enable us to decide which conclusion was correct, but my colleague Peter Berthold and I soon devised experiments that did. The bird we studied in the new experiments was the garden warbler (*Sylvia borin*), a species whose migratory patterns closely resemble those of the willow warbler. Three groups of garden warblers were exposed to a constant photoperiod for three years. For one group the period of light was 10 hours, for the second group it was 12 and for the third it was 16. In each case the results were the same: the migratory restlessness, molting and other related behaviors continued in a regular yearly pattern for the entire term of the experiment.

Subsequently we carried out an even longer experiment with garden warblers and blackcaps (*Sylvia atricapilla*). As before, the birds were exposed to an unchanging period of daylight. This time, however, we watched only for molting, which has turned out to be one of the most reliable indicators of cyclical yearly timing. Blackcaps and garden warblers in the wild molt twice a year. Under the experimental conditions both species continued to molt




**INTERNAL CUES underlie the timing of migration, as was suggested by an experiment carried out with three groups of willow warblers (*Phylloscopus trochilus*). Each pair of panels gives data for one group of birds. The upper curve indicates the photoperiod: the number of hours of daylight to which the birds were exposed. The lower curve shows the intensity of migratory restlessness (nocturnal activity displayed by caged birds when they would otherwise be migrating); the bar shows the timing of the winter molt, which is fixed in relation to migration. One group of warblers was held through the fall and winter in the author's laboratory in Germany and exposed to the normal variation in photoperiod there (*top*). Two groups of birds were denied that seasonal stimulus. One group was transported in September to equatorial Zaire, where there is little variation in photoperiod (*middle*). The other was held in the laboratory, where from September on it was exposed to a constant photoperiod of 12 hours (*bottom*). All three groups molted and underwent migratory restlessness on the normal yearly schedule, indicating that those events are motivated by an internal timekeeping mechanism rather than by seasonal cues such as the variation in photoperiod.**




**CIRCANNUAL RHYTHMS control the timing of seasonal events in some bird species, as was shown in a long-running experiment with blackcaps and garden warblers. Both species molt twice a year, once in winter and once in summer. The birds were held under conditions of constant temperature and photoperiod while the timing of the molts was observed. Results for one blackcap are shown in the upper panel, those for one garden warbler in the lower panel. Black bars correspond to summer molts, colored bars to winter molts. The molts took place on a regular schedule for up to 10 years under constant environmental conditions. The period between annual events, however, was only about 10 months. (For example, each pair of winter molts occurred about 10 months apart.) Hence the birds molted somewhat earlier each year. Because the period of such internal rhythms is not precisely one year, the rhythms are called circannual, from the Latin for "about a year."**

about twice a year for eight years, or even longer. The results of the two sets of experiments showed unambiguously that the small birds have a rhythmic internal timekeeper that functions to elicit the seasonal events in the right order.

It was notable that under constant environmental conditions the period of the internal rhythm was not exactly a year. For example, the winter molts of the caged blackcaps took place about every 10 months rather than every 12. This is the reason such endogenous oscillations are described as circannual, a word that comes from the Latin for "about a year." Circannual rhythms were first described some 25 years ago by Eric T. Pengelley and Kenneth C. Fisher of the University of Toronto in ground squirrels [see "Annual Biological Clocks," by Eric T. Pengelley and Sally J. Asmundson; SCIENTIFIC AMERICAN, April, 1971]. Since then similar periodic phenomena have been described in a broad range of organisms including coelenterates, insects, fishes, amphibians, reptiles and mammals in addition to birds.

The fact that the oscillations we observed in the garden warblers and blackcaps deviated from one year had two significant implications. First, it proved that the circannual rhythms could not have been due to the influence of uncontrolled seasonal stimuli present in the laboratory. If an uncontrolled seasonal stimulus had caused the observed rhythms, they would have had a period of precisely one year; that the circannual rhythms deviated from one year showed they are internal and spontaneous.

The second implication is that since in nature seasonal activities such as molting and migration always occur at the same time of year, there must be a factor that brings the approximately annual internal rhythms into correspondence with the solar year. The synchronizer was not difficult to find. We were able to synchronize a group of warblers so that they underwent migratory restlessness and molted on an exact annual cycle simply by simulating the yearly variation in photoperiod at 40 degrees north latitude, which lies in the warblers' European breeding ground. Even more striking was the finding that the circannual rhythm could be compressed to half its normal duration by increasing and decreasing the photoperiod on a six-month cycle instead of a 12-month cycle.

So far we had established that the circannual rhythm, modified appropriately by environmental factors, sets the pace for the overall seasonal pattern of events. It seemed important to find out whether endogenous rhythms can also affect the details of the migratory flight, such as its duration and even its ultimate target. In order to study those questions we had to work with birds that we knew could not have learned a specific migratory route from experience.

Now, almost all the birds in our experiments are reared by human beings. When the birds are about a week old, they are separated from their parents and raised by the laboratory staff, who feed them worms, insect larvae and other tempting morsels. This process, which we call hand raising, has at least two appreciable advantages for our kind of work. One is that by the time the birds are put in an experimental setup they are completely accustomed to the cage; hence they refrain from the erratic behavior seen in wild birds when first caged. Such erratic behavior would interfere with the experimental results, particularly in the study of migratory restlessness.

The other great advantage is that if the birds are hand-raised, their age, origin and history are known. As a result, when we came to study the influence of circannual rhythms on the details of the migratory pattern, we could rule out the influence of learning by working with birds we knew had never migrated. We found the inexperienced birds showed patterns of restlessness that mirrored the course of the migrational flight in wild birds. Caged willow warblers showed the most intense restlessness in August and September, which is the time free-living birds cross the Mediterranean Sea and the Sahara. Thereafter the nocturnal activity of the caged birds decreased, just as the migrational flight decelerates when the wild birds approach their winter home. The restlessness ended at about the time the free-living warblers reach the wintering ground.

Such results suggested that the willow warbler does not acquire the overall time course of its migration by learning. Additional work made it possible to generalize the conclusion. We found that in inexperienced, caged birds the duration of migratory restlessness was proportional to the distance covered by the species in its fall migratory flight. My colleagues Peter Berthold and Ulrich Querner showed that this correlation also holds within one species. They compared blackcaps from breeding grounds in Finland, Germany, France and the Canary Islands. The Finnish birds, which must travel farthest to reach central Africa, showed the most activity, followed respectively by the German, French and Canary Islands birds, which travel progressively shorter distances.

Berthold and Querner exploited the variation among the subpopulations of blackcaps to show that the time course of migration is genetically programmed. Their strategy was to cross German blackcaps with those from the Canary Islands to see what chronological pattern of migratory restlessness the offspring would display. If, as they hypothesized, the pattern was genetic, then the offspring should show a degree of restlessness intermediate between the two parental groups. The reason is that any behavior as complex as migration probably results from the operation of many genes, and the inheritance of polygenic traits often takes the form of an average of parental extremes.

Breeding blackcaps in captivity is an arduous and frustrating task, but Berthold and Querner persisted until they had obtained 32 hybrids from German and Canary Islands parents. The birds were hand-raised and held in cages until their first autumn, when the intensity of their migratory restlessness was measured. The results were just as predicted by the polygenic hypothesis: the hybrid birds developed a pattern of nocturnal activity almost exactly intermediate between those of the parental populations.

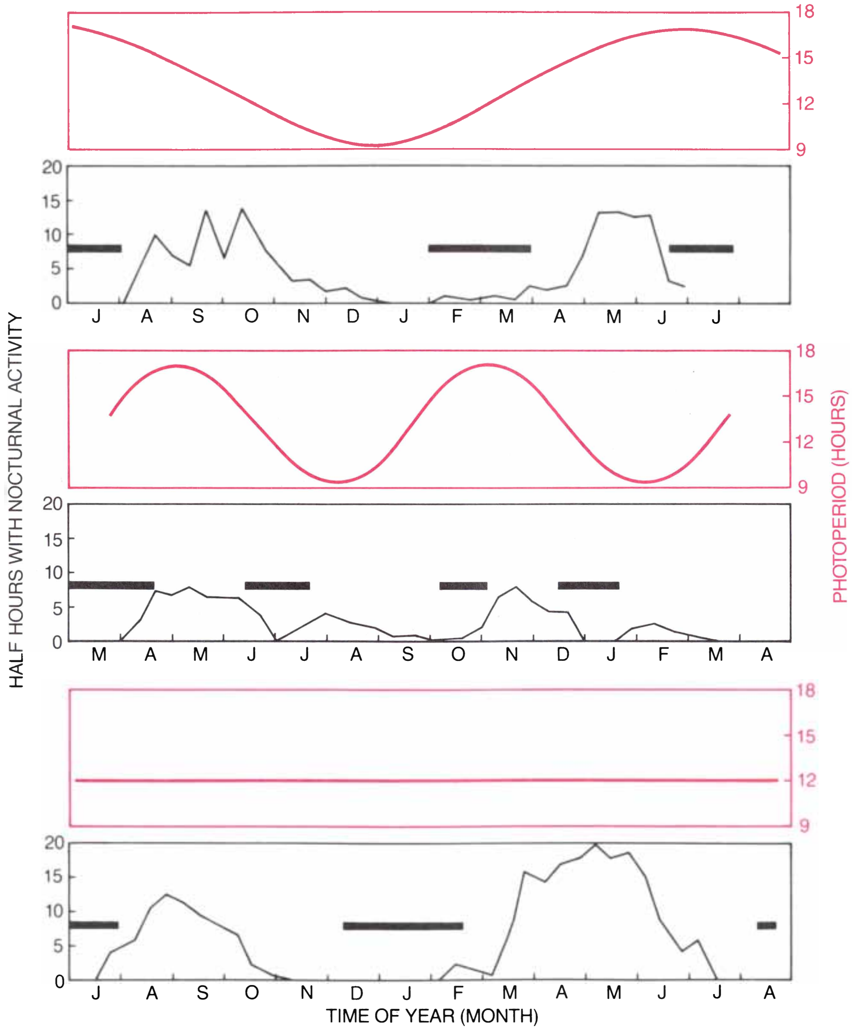
The breeding experiment carried out by Berthold and Querner showed indisputably that the temporal course of the migratory restlessness is determined by a genetic program. It soon became clear that this temporal program may also help young birds to find the wintering ground, which is clearly essential for survival.

Among species that migrate in groups, inexperienced migrants can follow more experienced leaders. Many species, however, migrate singly, including most of the birds we work with. Among those species each bird must be able to find the wintering grounds on its own, whether it has ever made the journey before or not. Clearly, if a bird has not migrated before, learning cannot help it to find its way, and so some other mechanism must serve for navigation. Evidence is accumulating that a temporal program is one such mechanism.

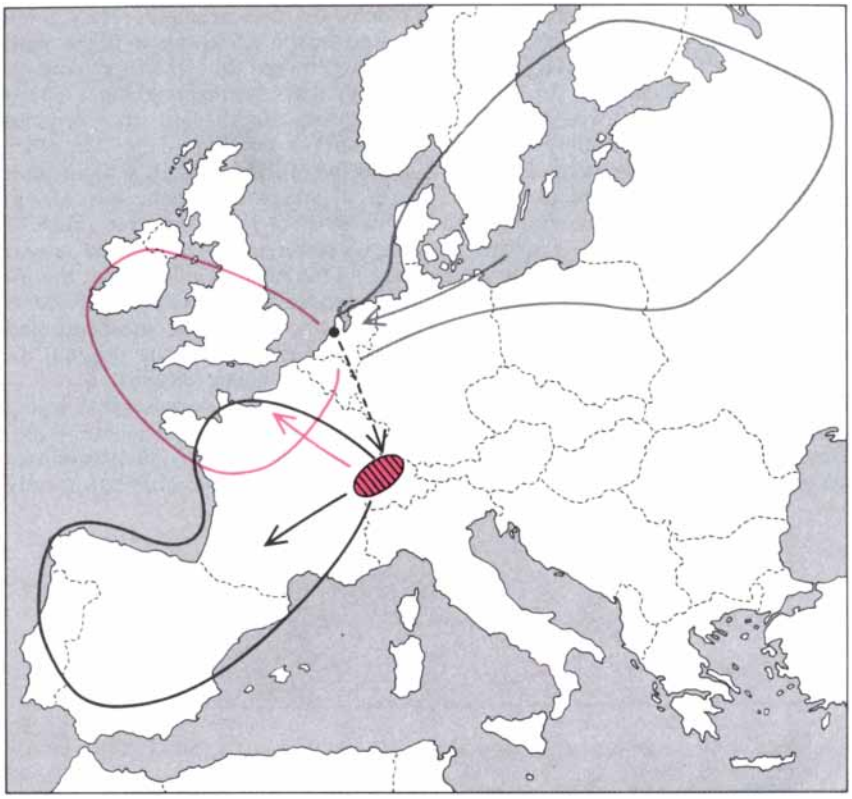
It has been known for some time that novices rely on methods of navigation different from those of experienced migrants. This was made strikingly clear in work done by A. C. Perdeck of the Institute for Ecological Research in the Netherlands. He studied European starlings, which breed on the shores of the Baltic Sea in Sweden, Finland, Latvia, Lithuania, Poland, Germany and Denmark. In the fall the starlings fly southwest to wintering grounds that are mainly in northern France and southern England. For one experiment Perdeck trapped 11,000 starlings near The Hague and had them transported by airplane to Switzerland, where they were banded and released.

More than 300 of the birds were later recovered, and when the recovery points were plotted on a map, an intriguing distribution emerged. Birds that had migrated at least once before were able to compensate for the southward displacement. They were found scattered along a flight path leading toward the wintering grounds, a path that entailed making a sharp northward turn from the original course. The young birds, on the other hand, were found along a flight path more or less parallel to the normal one, leaving them hopelessly far south of the usual target [*see illustration on next page*]. One must conclude that the inexperienced birds were unable to compensate for the displacement and had simply kept flying in the original direction after being released.

Perdeck's result suggests that young birds can fly in only one general direction when migrating. A navigation system capable of accommodating only




**SYNCHRONIZATION OF CIRCANNUAL RHYTHMS is accomplished by external stimuli such as changes in photoperiod. Because the circannual rhythms have a period that deviates from one year, they must be synchronized so that the physiological events underlying migration coincide with the seasons. The author showed that endogenous rhythms can be altered by variation in the photoperiod. One group of garden warblers was exposed to the seasonal changes of photoperiod in Germany; they underwent migratory restlessness and molted on the normal, annual schedule (*top*). A second group was held under constant conditions; they showed a circannual rhythm with a period of more than 12 months (*bottom*). A third group was exposed to a photoperiod that varied on a six-month cycle; their schedule for molting and migratory restlessness was compressed into six months (*middle*).**



**NAVIGATION SYSTEM of inexperienced birds differs from the one used by adults that have made the migratory flight at least once before. The difference was demonstrated by A. C. Perdeck of the Institute for Ecological Research in the Netherlands. Perdeck captured starlings about halfway along their fall migratory route (*gray*) from Baltic breeding areas to wintering grounds in France and England. The birds were transported to Switzerland and released. Experienced migrants (*color*) compensated for the displacement and flew toward the normal winter range. Inexperienced birds (*black*) kept flying in the original direction, on a course that took them toward the Iberian peninsula. It seems inexperienced birds are not capable of true navigation; they can fly only in one fixed compass direction.**

one flight heading is primitive, but it is by no means completely impractical. If the initial heading is correct and the external perturbations (such as those caused by wind) are not too great, all the flier needs to reach the target is a clock that indicates when to start flying and when to stop.

Various field experiments have yielded results consistent with the hypothesis that inexperienced birds depend on a navigational clock. Perhaps the clearest demonstration was made by W. Rüppell and E. Schüz of the Vogelwarte Rossitten, the predecessor of our institute. They trapped young carrion crows about halfway along their migratory route from Baltic breeding areas to wintering areas in northern Germany. The birds were banded and taken to a location beyond the normal winter range before being released. When some of the banded crows were recaptured during the winter, it was found they had continued to migrate on the original heading, which now took them away from the wintering ground. Moreover, they had flown a distance roughly equal to the distance that separated them from the target when they were captured. It would seem that the migratory clock had simply continued to tick after the displacement, with the result that the young crows had flown the correct distance, but away from the target.

Perdeck's results, combined with those of Rüppell and Schüz, help to clarify the difference between the navigation systems of experienced and inexperienced birds. Birds that have migrated before have the capacity for true navigation: when displaced, they can correct their course and find the target. (How they accomplish the feat is not yet understood.) Inexperienced birds have no such capacity; they apparently rely solely on a single heading and a preset flying clock. Until recently the nature of the navigational timekeeper was not known. Our work with warblers, however, suggests the clock is provided by the genetic program that specifies the temporal course of migratory restlessness. The physiological rhythm underlying the restless behavior seen in the caged birds tells the migrant when to take off, when to expend peak effort, when to decelerate and when to stop flying.

Although their navigation systems differ considerably, novices and experienced birds share the need to know the correct flight heading. Intriguingly, it appears circannual rhythms can help to solve that problem too. For some time experimental results have been available implying that the migrant's choice of a compass heading is determined by its physiological state rather than by external cues such as geographic information or the position of the stars. Among the data are some obtained by Stephen T. Emlen of Cornell University [see "The Stellar-Orientation System of a Migratory Bird," by Stephen T. Emlen; SCIENTIFIC AMERICAN, August, 1975].

Emlen exposed a group of indigo buntings to two photoperiodic cycles a year. The accelerated changes in photoperiod advanced the timing of molt and of migratory restlessness in the experimental birds. Eventually the birds came into an "autumnal" physiological state in the spring. They were then put in orientation cages in a planetarium under a simulated spring sky. In spite of the celestial cues, the birds responded to their internal state and consistently hopped southward, as if attempting to begin the fall migratory flight.

From the point of view of the work my colleagues and I have been doing, Emlen's results could be interpreted in two ways. The buntings' choice of the incorrect direction could be due either to the effect of the photoperiod or to that of the underlying circannual rhythm. We needed to know which factor was responsible. Therefore, with Wolfgang Wiltschko of the University of Frankfurt as my collaborator, I set out to design experiments that would disentangle the two effects. Wiltschko and I did our experiments on 59 hand-raised garden warblers. Garden warblers are among a large group of European birds that share a basic migratory route. They leave their breeding ground in the fall on a southwesterly course. Over the Iberian peninsula or northern Africa, however, they turn south or south-southeast and continue on that heading until they reach their winter range. The return flight in spring has a more direct route: the birds fly straight from south to north.

As in our previous work, to isolate the effect of the circannual rhythm we needed to minimize the seasonal fluctuations that the birds experienced. The warblers were held in chambers with constant temperature and an unvarying photoperiod.

In order to test a bird's choice of direction, the bird must be able to orient itself in relation to the points of the compass. Wiltschko had shown previously that the garden warbler can orient itself by means of either the stars or the earth's magnetic field. The celestial map changes with the season. Since we intended to eliminate all seasonal cues, the birds were not allowed to see the sky. The earth's magnetic field, on the other hand, undergoes very little seasonal variation. We made no effort to insulate the birds from the magnetic field, and they were thereby able to orient themselves.

The experimental system was nearly complete. The warblers had been isolated from seasonal cues and provided with a means of sensing compass directions. Now we needed a way to test the caged birds' directional preferences during the migratory period. In order to observe their directional preferences the warblers were put in special cages at regular intervals throughout the year. Each cage was an octagon with a diameter of about a meter. Eight perches radiated from the center of the cage like the spokes of a wheel. Microswitches mounted under the perches made it possible for us to tell which perches the birds preferred during the night in each phase of the migratory cycle.

The results were quite dramatic. In August and September, when wild garden warblers head southwest on the first leg of their triangular flight path, the caged birds tended to hop onto the perches at the southwest corner of the cage. During the second half of the fall migratory season, when the migrating birds turn south and head for central Africa, the caged birds preferred to perch in the south and southeast parts of the cage. In the spring, when the migrants head straight for Europe from their equatorial winter homes, the caged birds most often chose the north perches.

It seems clear that circannual rhythms can control the direction of migration as well as its initiation and its temporal structure. Indeed, the overall result of the work done in my laboratory in the past two decades has been to emphasize the effects of the spontaneous internal clock. Yet it should not be supposed that the operation of the circannual rhythms eliminates the influence of external cues. As I have described above, circannual rhythms can be modified by changes in photoperiod or—in adult birds—by navigational learning. Some recent work has been aimed at finding out more about modifications of the internal program; my colleague Herbert

Biebach has shown that the modifications can be quite specific.

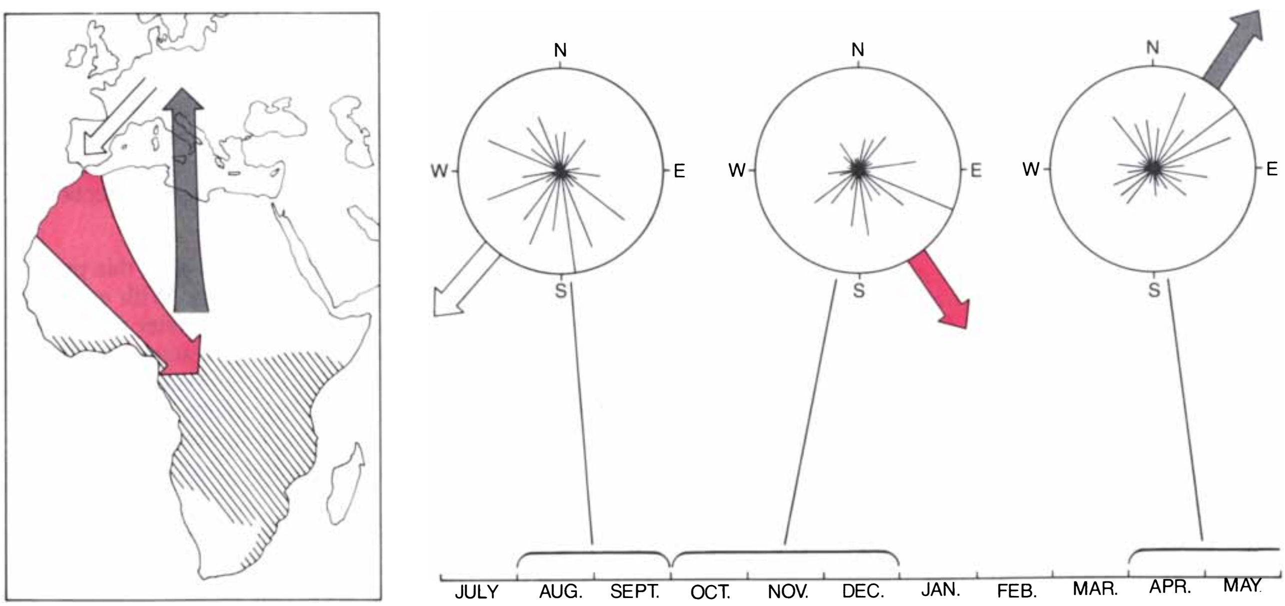
Biebach concentrated on the changes in weight that birds undergo while they are migrating. Before taking off, the migrants generally eat more intensively than usual in an attempt to acquire stores of fat to be used as fuel on the journey. The weight gain can be considerable. Garden warblers in the field can grow from an initial weight of 16 grams to a final weight of 30 grams, and in the laboratory (where an unlimited supply of food is available) the gain can be even greater. The accumulated fat stores dwindle rapidly as the birds make the exhausting migratory flight.

Biebach trapped migrating songbirds at two sites in the Sahara. One site was an artificial resting area made of a few bushes planted in the sand. Birds were attracted by the bushes and touched down, but no food or water was available there. The other site was an oasis with abundant vegetation and a rich insect life. At both sites many of the birds present in the morning were trapped, banded, weighed and then released. Some birds were later recaptured and weighed again. It turned out that the birds behaved quite differently at the two sites. At the desert trap, where there was no food, all the birds left on the evening of the day they arrived. At the oasis, however, a fairly large proportion of the arriving birds stayed for more than one day. The weight data showed that the birds remaining for more than a day tended to be the ones whose fat stores had been depleted on arrival.

Biebach's findings suggested that the combination of low fat reserves and the opportunity to replenish them inhibits the endogenous migratory program, which would otherwise impel the birds to continue their flight. That conclusion has since been confirmed in the laboratory. The laboratory studies show that neither the depletion of the fat reserves nor the opportunity to feed is on its own sufficient to interrupt the operation of the program; both factors must be present simultaneously for the migratory impulse to be inhibited. As soon as the bird attains a certain minimum body weight the impulse reappears.

The capacity of the endogenous program to be overridden by external stimuli has decided advantages. If the circannual rhythms could not be modified, a bird whose fat reserves were depleted might starve to death while being driven on by the relentless migratory impulse. If the impulse can be temporarily overridden, the migrating bird has a chance to replenish its stores of fat before resuming its flight. In one sense the behavior Biebach has illuminated so nicely is similar to the navigation system of the experienced birds. In both instances a rather inflexible mechanism is overridden by a system that can adjust the internal time program to specific external conditions.

One of the main tasks facing students of bird migration is to find out just how the endogenous program interacts with external stimuli to yield the observed migratory behavior. Another problem, equally fundamental, is to identify the physiological basis of the circannual rhythms. My colleagues and I are currently working on both questions, but it may take many years to answer them. After all, it has taken two full decades to achieve a detailed understanding of how the circannual rhythms operate. In those two decades it has become clear that endogenous rhythms provide the overall framework of migratory behavior. The details of the migratory cycle can be modified in a variety of ways to accommodate external circumstances, but the impulse that drives the great V of geese as well as the ragtag group of starlings comes from yearly rhythms originating deep in the organism.




**DIRECTIONAL PREFERENCES of garden warblers are controlled by the circannual rhythms, as was shown in an experiment carried out by the author and Wolfgang Wiltschko of the University of Frankfurt. The garden warbler migrates on a triangular course (*map at left*). In September it leaves Europe on a southwest heading. In late fall it changes course over the Iberian peninsula and heads southeast to equatorial Africa. The return journey in the spring takes a direct south-north route. The three circles at the right show the directional preferences of garden warblers held under constant conditions of light and temperature in orientation cages like the one on page 85. Each circle corresponds to the results for a particular time of year. The lines within the circles show how often the warblers preferred particular compass directions; each line represents 15 degrees of the compass. The direction most often chosen is shown as a radius of the circle, and all other lines are proportional to that one. The large arrow projecting from the edge of the circle shows the overall direction that results from summing the choices of the caged birds. In August and September the birds preferred a southwesterly direction, in October, November and December a southeasterly direction and in April, May and June a northeasterly one. All three results correspond approximately to the migratory directions of the free-living birds at those times of year.**

# Optical Gyroscopes

*They can detect rotational motion just as mechanical gyroscopes can, but they contain virtually no moving parts. The instruments have already established an important place in guidance systems*

by Dana Z. Anderson

The word gyroscope brings to mind a picture of the string-powered novelty capable of balancing itself on the tip of a pencil. Such a toy is actually a primitive representative of a family of instruments whose essence is a framework containing a rapidly spinning wheel. The angular momentum of the wheel causes it to resist change in its orientation even as the framework is rotated. A gyroscope can therefore reveal the extent of the rotation; in this way it can also provide directional information for navigation. Indeed, nearly all vehicles more sophisticated than an automobile (and some that are much less sophisticated) depend on gyroscopes to keep them on course. Gyroscopes are at the heart of the inertial-guidance systems of any jet airliner or oceangoing ship, for example. The reason is that the instruments require no external stimulus in order to function. A gyroscope can, for instance, detect motion even within the confines of an isolating box that prevents a voyager from observing the stars and deprives him or her of a magnetic field for a compass.

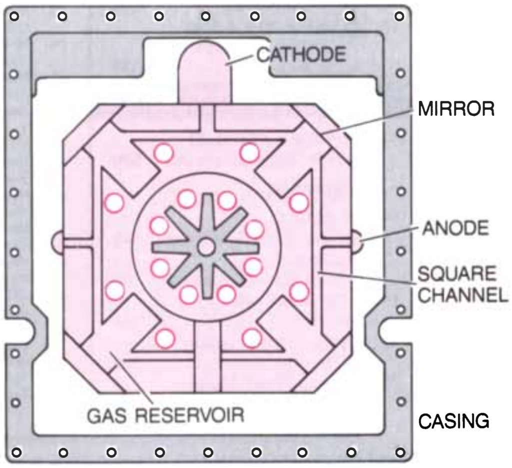
At about the turn of the century investigators realized that light could also exhibit gyroscopic behavior: the amount of time required by light to traverse a circular pathway depends on whether the pathway is stationary or rotating. The time difference can serve as a measure of the amount of rotation. The development of a practical optical gyroscope, however, had to await the advent of the laser and advances in such optical technology as fiber-optic cables and highly reflective mirrors. From this technology there have emerged two classes of optical rotation sensors: fiber gyroscopes and ring-laser gyroscopes.

The ring-laser gyroscope, technologically the more mature of the two, is a fascinating instrument, and it has already established itself as a practical element in guidance systems. New passenger carriers such as Boeing's 757/767 series and a number of Airbus Industrie's A310's rely on ring-laser gyroscopes rather than mechanical ones. Although the sensitivity of the new instruments is extraordinary, it is not for want of accuracy that the niche of the optical gyroscope in navigation is widening. Mechanical gyroscopes are accurate too, but their moving parts make them more complex than the optical devices, which have virtually none. In spite of the fact that optical gyroscopes contain some moving parts, they are still easier to maintain, and therefore less costly, than their mechanical counterparts.

To understand how light can be manipulated to measure rotation, imagine a circular ring with an observer fixed to it. Suppose the observer emits a pulse of light in such a way that half of it traverses the circle in each direction. If the ring is not rotating, clearly both halves of the pulse will return to their point of origin (the observer) at the same time. If, however, the ring is rotating, say, counterclockwise, the point of origin will be moving toward the clockwise-traveling pulse and away from the counterclockwise-traveling pulse. The observer will therefore encounter the clockwise-traveling pulse before the counterclockwise one. The difference in arrival time is directly proportional to the ring's rotation; it does not matter whether or not the ring rotates about its center.

The difference in round-trip travel time is attributed to a difference in path length. In the example given above, the path traversed by the counterclockwise pulse is said to be longer than the path traversed by the clockwise pulse. The rotation-induced difference in path length is called the Sagnac effect, after Georges Marc Marie Sagnac, who first demonstrated the phenomenon in 1913.

The Sagnac effect is most readily apparent in the fiber gyroscope. I shall therefore discuss this gyroscope first, even though its development actually began 12 years after that of the ring-laser gyroscope, and fiber-gyroscope technology has not yet produced a truly practical rotation-sensing instrument. (Success does, however, appear to be in the offing.) The heart of the fiber gyroscope is a long optical fiber, typically from half a kilometer to one kilometer long, that is coiled for com-

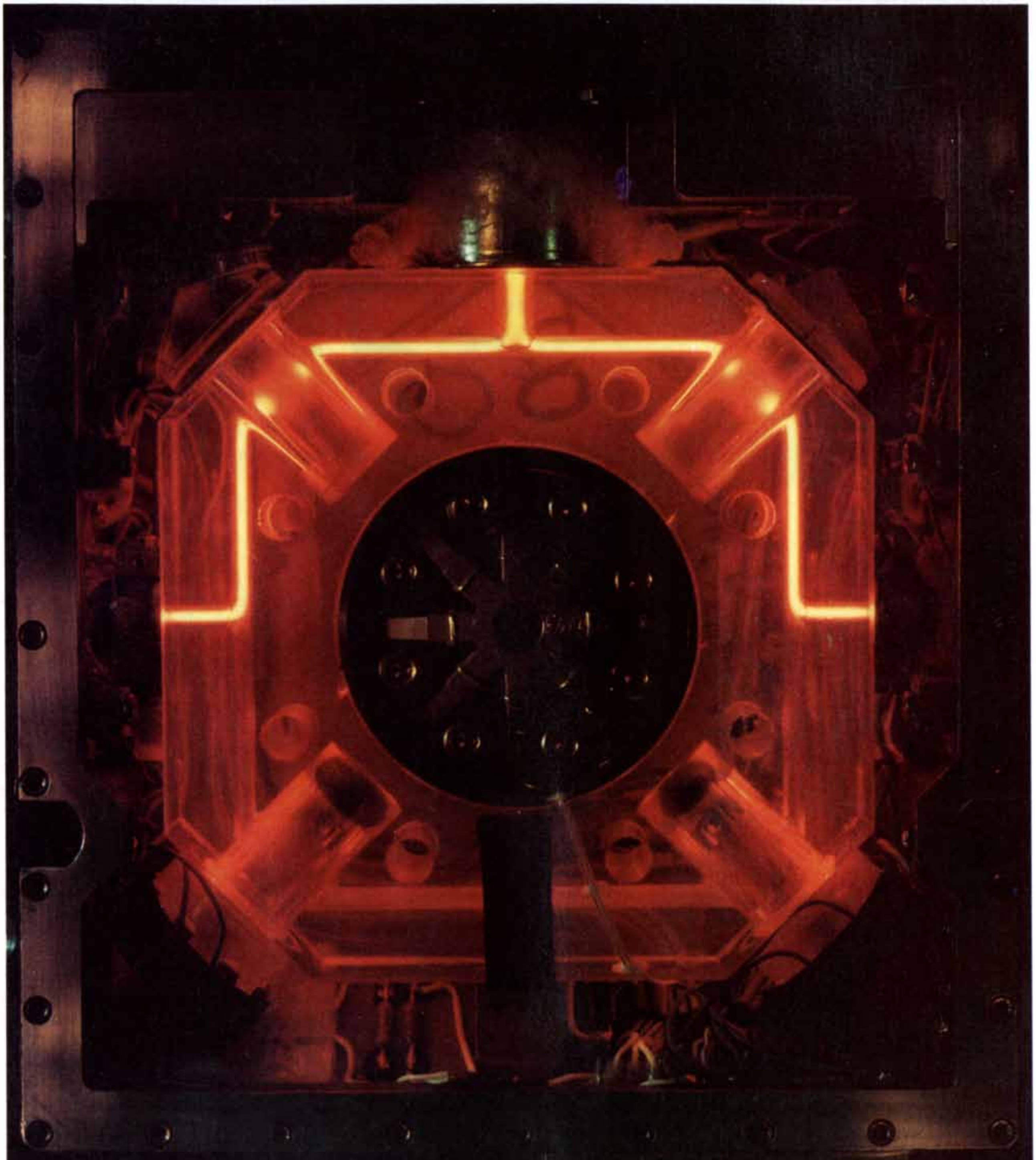



**RING-LASER GYROSCOPE is typically made from a single block of glass, as is seen in the photograph on the opposite page. The square channel drilled in the block contains a mixture of gases. A flow of electric current between the cathode, or negative electrode, and the two anodes, or positive electrodes (*see map above*), excites the atoms of the gas and causes them to emit light. Mirrors at the four corners of the block limit the light to travel within the square channel, resulting in the formation of a standing wave. (The glow in the top half of the square channel is from a phenomenon known as plasma discharge; the standing wave itself cannot be seen.) The black tape at the bottom of the photograph covers details that the manufacturer, Litton Guidance and Control Systems, considers to be proprietary information.**

pactness. The fiber acts as a light pipe, confining light within its walls. The reason for working with such a long fiber is that differences in path length, and therefore in travel time, increase with increasing fiber length, facilitating rotation measurements.

The fiber gyroscope exploits the extremely small wavelength of light to measure accurately the path-length difference between the distances traveled by two counterpropagating light beams. To generate the beams a semitransparent mirror, or beam splitter, splits a beam from a light source (often but not necessarily a laser) in half. The two beams travel in opposite directions through the coiled fiber ring and emerge to be recombined at the semitransparent mirror. If a screen is placed at the output of the mirror, no light appears on it when the fiber coil is stationary. The reason is that the two beams travel identical path lengths and the semitransparent mirror introduces a "phase shift": the beam reflected by the mirror is precisely out of phase with the one transmitted by the mirror. At the mirror output the two beams therefore interfere destructively: they cancel each other exactly.

What happens when the fiber coil rotates? (By this I mean that the coil, light source, semitransparent mirror and screen rotate together; the coil does not spin with respect to the other



components.) Suppose, for instance, the components are mounted in an airplane and the craft suddenly veers. In this case the two counterpropagating beams in the fiber travel slightly different path lengths. At the output of the semitransparent mirror they no longer cancel each other exactly, and so a bright spot appears on the screen. If the coil (and the other components) rotates fast enough so that the difference in path length is half the wavelength of the light, the spot becomes as intense as the original light source. I have described a bare-bones version of the fiber gyroscope. In a practical device sophisticated electronic techniques are employed to monitor the output of the semitransparent mirror and thereby obtain the rate of rotation.

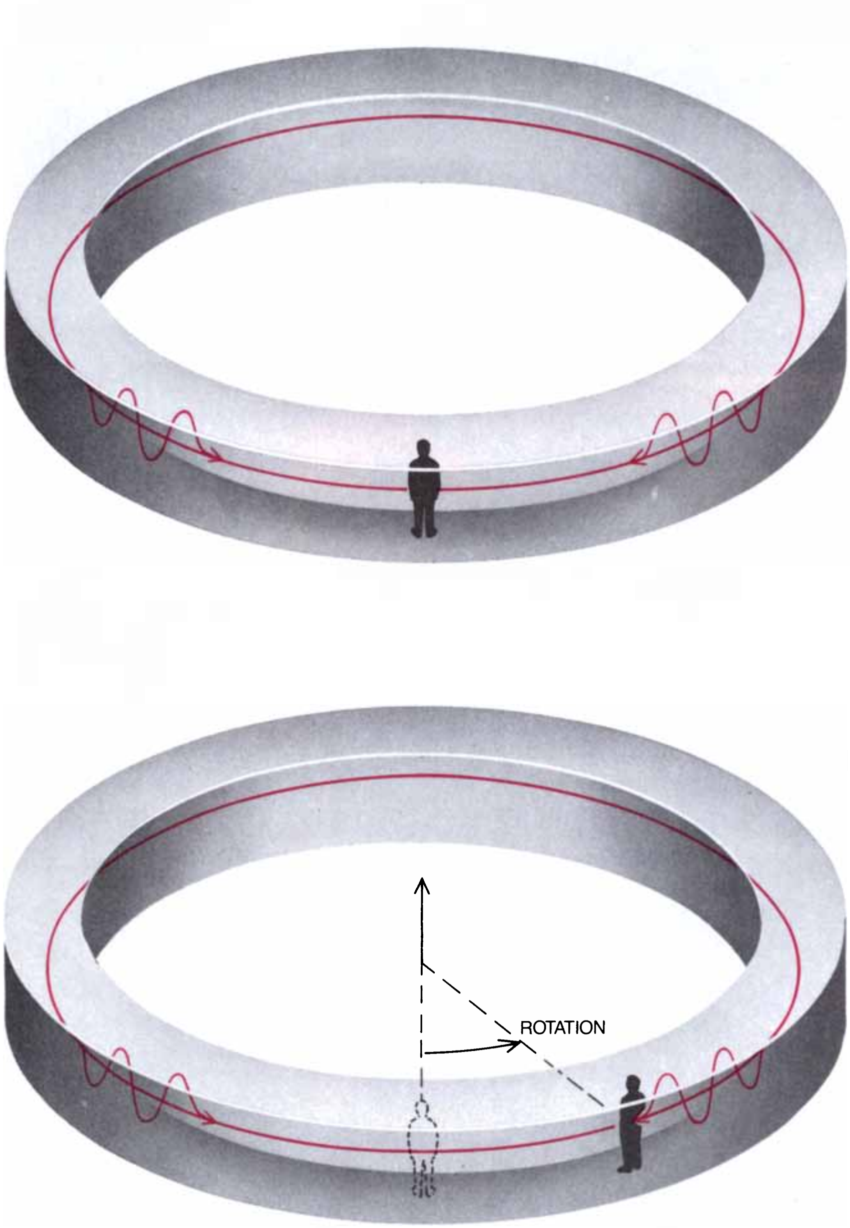
Work on the more established ring-laser gyroscope dates back to shortly after the publication in 1958 of a paper by Arthur L. Schawlow and Charles H. Townes setting forth the basic principle of the laser. Clifford V. Heer of Ohio State University soon realized that a resonant cavity could be exploited to measure rotation rates. A resonant cavity is a hollow cavity for reinforcing a sound wave or an electromagnetic wave. A laser is in essence nothing more than a resonant cavity. It typically consists of a long, straight tube filled with an amplifying medium, a solid, a liquid or a gas. Each end is capped with a highly polished semitransparent mirror. As light travels back and forth between the two mirrors its intensity is amplified. Laser output takes place when some of the light passes through one of the mirrors.

Heer recognized that by forming a ring-shaped resonant cavity one could create an optical gyroscope. In such a gyroscope light circulates many times around a given path rather than just back and forth between two mirrors. Both Heer and, independently, Adolph H. Rosenthal, then at the Kollsman Instrument Corporation, made subsequent advances in the theory. In 1963 Warren M. Macek and Daniel T. M. Davis, Jr., of the Sperry Gyroscope Company demonstrated the first ring-laser gyroscope. Their instrument consisted of four glass tubes, each one meter long, arranged in a square. They made light travel around the device by placing a mirror at each corner.

Technological progress in the ensuing years has been astonishing: a sensitive ring-laser gyroscope can now fit in the palm of the hand. A typical gyroscope is made from a single block of glass into which a square channel is drilled. A mixture of gases, such as helium and neon, fills the channel. To complete the laser a small number of electrodes and four mirrors are attached to the block. Four is not a magic number; some ring-laser gyroscopes have a triangular channel and three mirrors, others have a hexagonal channel and six mirrors.

Even though a practical ring-laser gyroscope has the shape of a polygon, its operation is most easily understood by imagining the ideal case of a circular ring. When the ring is stationary, a discharge in the lasing gas generates a standing light wave within the ring. As in the case of a vibrating violin string, along the wave there is an alternating series of nodes ("valleys," or stationary points) and antinodes ("peaks," or points of maximum oscillation). Because the wavelength of light is extremely small—about .6 micrometer (a micrometer is a millionth of a meter)—a large number of nodes and antinodes fit around a ring. The wave in a ring with a 30-centimeter perimeter, for instance, has on the order of a million nodes and antinodes.

Now suppose the ring rotates, that is, the fictitious airplane carrying the gyroscope veers. What happens? The standing wave remains fixed in an in-

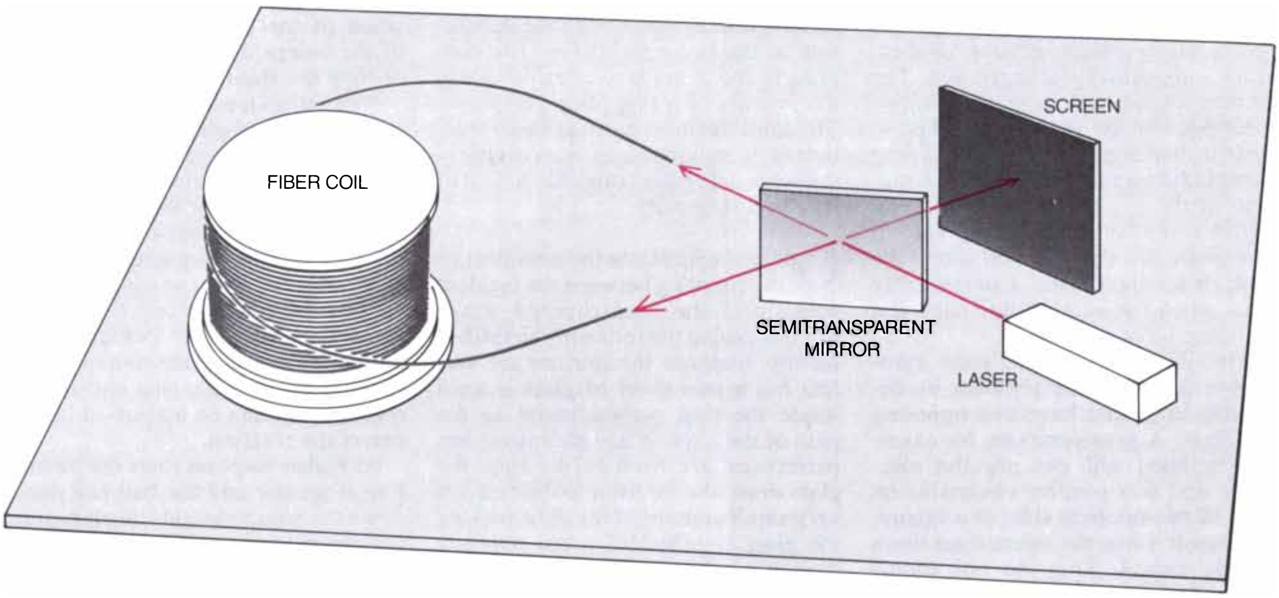


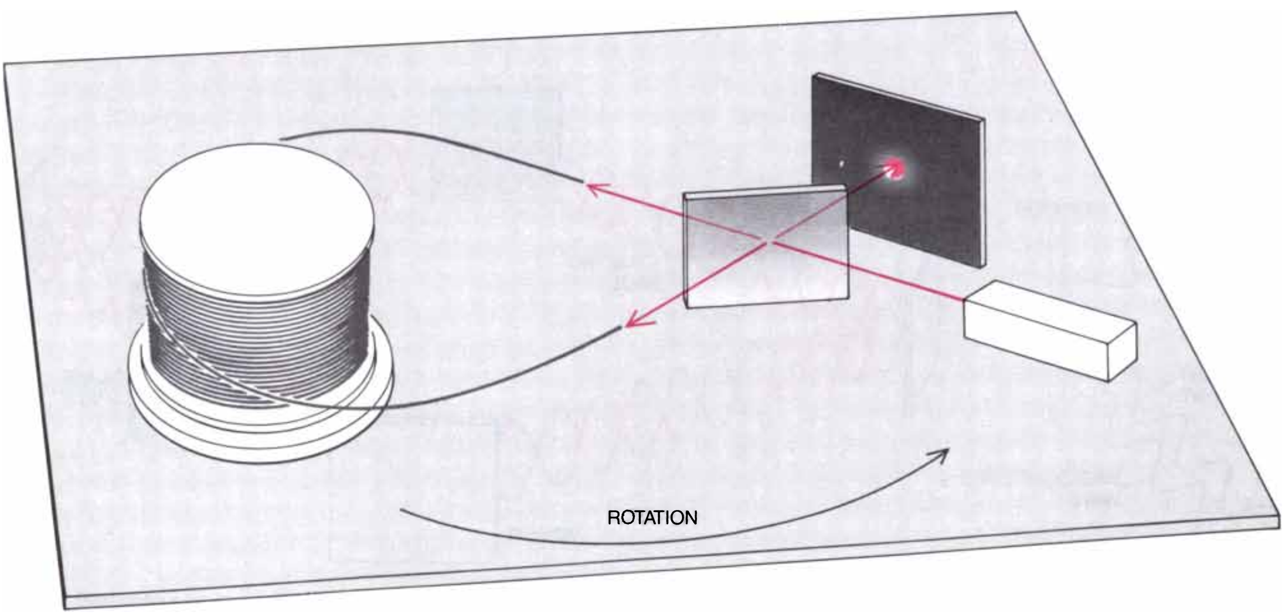

**BASIC PRINCIPLE underlying the optical gyroscope is that the amount of time required by light to traverse a circular pathway depends on whether the pathway is stationary or rotating. Suppose a hypothetical observer in a stationary ring emits a pulse of light and two halves of the pulse travel in opposite directions around the ring (*top*). The observer receives both pulses simultaneously. If the ring is rotating, however, the observer moves toward one of the pulses and away from the other, and the pulses are received at different times (*bottom*). The time difference is proportional to the rate at which the ring is rotating.**

ertial, or nonrotating, frame of reference; this is how the Sagnac effect manifests itself in the ring-laser gyroscope. An observer rotating with the ring would pass by the nodes and antinodes of the standing wave. The number of nodes the observer passes is directly proportional to the angle through which the ring has rotated. By counting the passage of the nodes one can determine how much the ring, and therefore the airplane, is rotating.

A noncircular ring (such as a triangle or a square) can also be employed to detect rotation. In this case, however, the standing wave does not remain fixed in an inertial frame of reference when the ring turns; instead it rotates at a lower rate than the ring does. The rotation rate of the standing wave depends on the shape of the ring. As far as the rotating observer is concerned, however, his or her rotation rate is still directly proportional to the rate at which nodes pass a given point on the ring.

If the ring-laser gyroscope were an ideal device, the rate at which nodes pass a given point on the ring would be directly proportional to the rate of rotation of the ring. If the ring were not rotating, for instance, the position of the nodes would remain fixed. In practice, however, two major sources of error cause the behavior to deviate from




**FIBER GYROSCOPE can serve as a measure of rotation by detecting interference between two light beams traveling in opposite directions around a long, coiled fiber. The beams are generated by splitting a laser beam with a semitransparent mirror. After traveling through the fiber the two beams recombine at the semitransparent mirror. When the apparatus is not rotating, the beams interfere destructively, so that there is no light output (*dark screen at top*). When the entire apparatus (including the laser, the semitransparent mirror, the fiber coil and the screen) is rotating, the two beams no longer cancel each other perfectly, and so there is a bright spot on the screen (*bottom*); the brightness, or intensity, of the spot increases with the rotation rate. The rotation rate of the apparatus, and hence of the carrier on which it is mounted, is inferred by monitoring the intensity of the spot. Fiber gyroscopes are still under development.**

the ideal. Fortunately neither is fatal, for they can both be remedied.

The first source is a bias effect that makes the standing wave rotate even when the ring is stationary. Bias arises when the gas inside the ring flows. Gas flow is an unwanted consequence of the fact that, in order to generate the standing light wave, power must be supplied to the ring. This is done by applying a large voltage difference between a positive electrode, or anode, at one site on the ring and a negative electrode, or cathode, at another site. The high voltage ionizes some of the gas and creates a plasma: a "soup" of electrons and positively charged ions. The electrons tend to drift to the positive electrode and the positive ions tend to drift to the negative electrode. These flows induce a rather complicated motion of the neutral atoms in the gas, which results in a net flow of neutral atoms around the ring and along the path of the light beam. Consequently even when the ring is stationary, the standing wave can rotate.

Manufacturers of ring-laser gyroscopes minimize the problem by designing rings that have two opposing gas flows. A gyroscope can, for example, be fitted with one negative electrode and two positive electrodes on each of two opposite sides of a square. The result is that the induced gas flows largely cancel. They do not cancel completely, but a stable residual bias can be measured and compensated for, and consequently only unpredictable changes are a source of bias error.

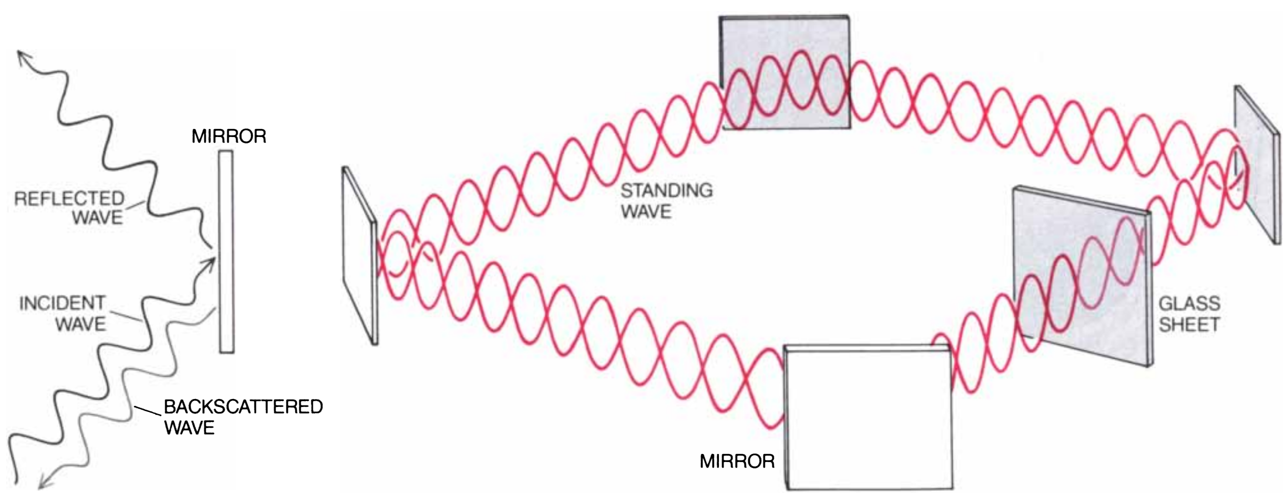
The second and more serious source of error in the operation of the instrument is a phenomenon called frequency locking. It causes the standing wave to "lock" onto the ring, so that an observer on the ring could not tell whether or not the ring is rotating. Frequency locking, whose effects are in general noticed only at relatively low rates of rotation, plays a role analogous to that of friction in a mechanical gyroscope. The standing light wave can be viewed as two waves traveling in opposite directions around the ring. Frequency locking arises from a coupling interaction of the two light waves. The coupling is due to minute imperfections in the mirrors of a ring-laser gyroscope. The imperfections cause a small fraction of an incident light wave to backscatter in a direction opposite to that of its initial trajectory.

One can appreciate the net effect of the coupling between the incident waves and the backscattered waves by considering the following simplified picture. Suppose the mirrors are perfect but a thin sheet of glass is fixed inside the ring perpendicular to the path of the light. (Since the mirror imperfections are fixed to the ring, the glass must also be fixed to the ring.) A very small amount of the light striking the glass is reflected in the opposite direction. The standing wave of the ring-laser gyroscope "prefers" to have a node at the sheet of glass. Since the glass is fixed to the ring, the standing wave will therefore "try" to rotate with the ring, and at low rotation rates it succeeds in doing so.

For a fruitful analogy to frequency locking imagine a ball, immersed in a viscous fluid such as maple syrup, rolling down a bumpy hill. The force of gravity represents the ring's rotation: a steeper hill corresponds to a higher rotation rate. The viscous fluid prevents the ball from accelerating indefinitely; instead it reaches a maximum speed. The bumps represent the nodes and antinodes of the standing wave, and the ball represents the position of the glass sheet. The height of the bumps depends on the amount of light the sheet reflects.

When there is no rotation, there is no incline and the ball rests in one of the local valleys between two bumps: it is at a node. Then the ring rotates slowly; by analogy, a slight incline develops. The ball rolls downward a little bit, but the nearest bump presents an upward slope that the ball cannot climb, and so the ball remains trapped. This corresponds to frequency locking, that is, a situation in which the standing wave remains fixed to the ring and the gyroscope registers no output—it fails to detect the rotation.

At higher rotation rates the inclination is greater and the ball can dance down the bumpy hillside. Such inclines correspond to rotation rates at which the frequencies unlock and the ring can once again sense rotation: the standing wave is no longer fixed to the ring and the gyroscope produces an output. As the rotation rate is in-



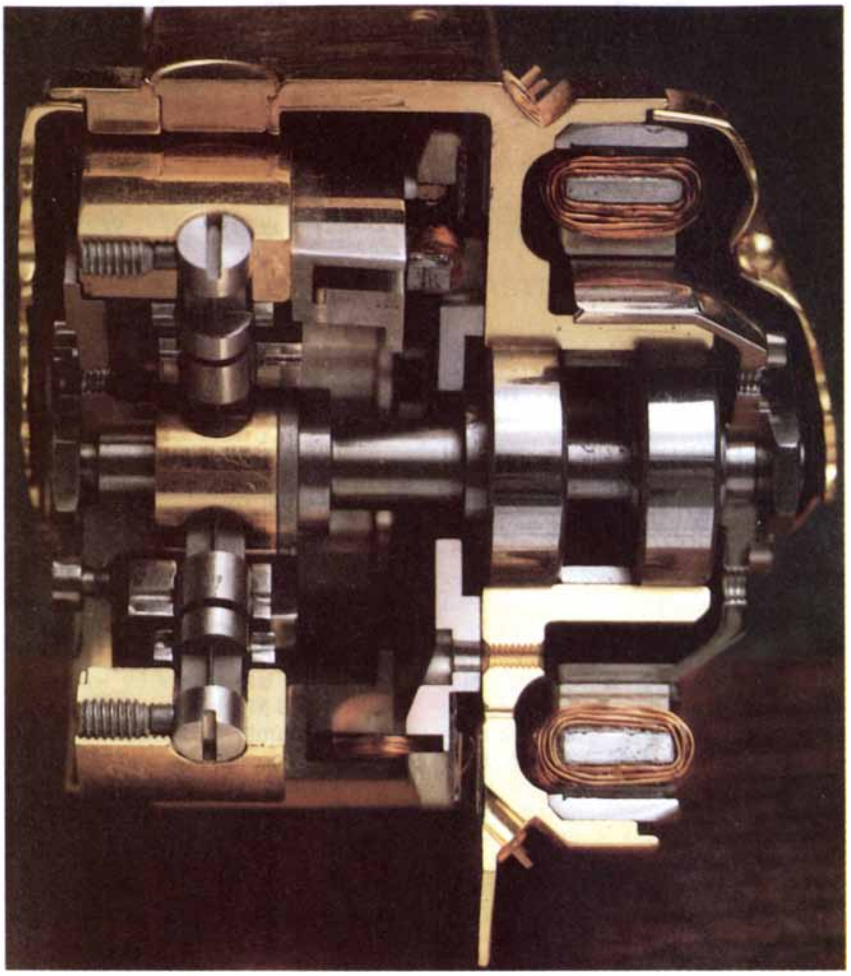

**FREQUENCY LOCKING is a potential source of error in the operation of a ring-laser gyroscope. It causes the standing wave in the instrument to "lock" onto the ring, so that an observer on the ring cannot tell whether or not the apparatus is rotating. The standing wave can be thought of as two waves traveling in opposite directions around the ring. The locking phenomenon arises from minute imperfections in the mirrors of the gyroscope. The imperfections cause a small fraction of an incident light wave to backscatter in a direction opposite to that of its original trajectory (*left*). Coupling between the waves results, the net effect of which can be understood by imagining that the mirrors are perfect but that a thin sheet of glass is fixed inside the ring perpendicular to the path of the light (*right*). The standing wave "prefers" to have a node (a "valley," or stationary point) at the position of the glass. Because the glass is fixed to the ring, the standing wave will "try" to rotate with the ring. At low rotation rates it succeeds, and frequency locking takes place. The effect can be minimized by oscillating the ring at a fairly high rate with a mechanical dither; the motion "shakes loose" the standing wave.**

creased still further the perturbation of the motion of the ball due to the presence of the bumps becomes even less significant.

Frequency locking at low rotation rates has bedeviled the optical-gyroscope industry since the beginning. Much effort has gone into eliminating the problem. The most successful solution so far has been a device known as a mechanical dither that rotates the ring back and forth at a fairly high rate. The general idea is to keep the ring moving so that frequency locking cannot take place. Since no net rotation is introduced by the oscillation, the dither has no net effect on measurements. It is something of a disappointment, however, to resort to this kind of solution, inasmuch as the ring-laser gyroscope's big advantage was supposed to be the lack of moving parts: although the dither movement is indeed slight, the dithering apparatus is regrettably complicated (albeit far less complicated than a mechanical gyroscope). Investigators still continue to search for a better solution.

The bias and the frequency-locking problems are technical ones; in principle they can be completely eliminated. Ultimately the sensitivity of the instrument must succumb to more imperative constraints: the dictates of quantum mechanics and the Heisenberg uncertainty principle. In its simplest form the uncertainty principle states that it is impossible to simultaneously know both the exact position and the exact velocity of a particle, such as an electron. Applied to the ring-laser gyroscope, it means the standing wave will not remain truly at rest even when the ring is stationary. Remarkably, technology has advanced to the point where ring-laser gyroscopes can perform to within better than a factor of 10 of the limits imposed by quantum mechanics.

The fundamental source of error in a ring-laser gyroscope is a manifestation of the uncertainty principle known as spontaneous emission. Power input excites the electrons of the atoms composing the amplifying medium within the optical ring. Lasing takes place when a photon, or quantum of light, stimulates an atom and causes an excited electron to make a transition to a lower (unexcited) state. When the transition occurs, the atom emits a photon that has the same direction, frequency and phase as the incident photon. Every so often, however, an electron in an atom will make a transition spontaneously, emitting a photon that has an arbitrary direction and phase. Some of these photons are emitted in the direction of the stimulated wave and change the net phase of the wave by a small random amount. In the course of time the spontaneous emissions make the standing wave drift from its initial position: the gyroscope develops an angular displacement error.



**MECHANICAL GYROSCOPE operates on the same principle as a toy gyroscope, but it is much more complicated. Its role in navigation is being challenged by the optical gyroscope.**

In general, high-accuracy applications require large gyroscopes because they minimize the relative effect of the displacement error. Mirror quality also affects accuracy. A perfect mirror would reflect all incident photons. In the real world, however, each time a photon hits a mirror there is a small chance that it will be scattered, absorbed or transmitted. The effectiveness of a gyroscope increases with increases in the number of round trips made by each photon before it is lost to absorption, transmission or scattering. For this reason the development of ring-laser gyroscopes has generated a dramatic advance in low-loss mirror technology. A good ring-laser gyroscope mirror loses less than one photon in 5,000 photons. (A bathroom mirror loses about one photon in 20.)

In spite of the apparent differences between fiber and ring-laser gyroscopes, nature conspires through quantum mechanics to make their theoretical performances similar. A fiber coil that has a given number of turns is fundamentally equivalent to a ring-laser gyroscope whose photons undergo the same number of round trips. Of course, the instruments must have the same dimensions and optical powers, and they must be operated with light of the same wavelength.

No optical gyroscope is ever likely to fascinate a child as much as the toy version of its mechanical counterpart. The success of a practical device, however, would seem to disregard any such considerations and ultimately become an issue of economics. To be sure, the driving force behind optical-gyroscope technology is one of cost. On the other hand, the beauty of the optical gyroscope's underlying principle and elegant construction is more than a mere reflection of its greatest asset as a navigational instrument.

# The "Stress" of Being Born

*The stress of journeying through the birth canal is not harmful to most infants. In fact, the surge of "stress" hormones it triggers can be important to the neonate's survival outside the womb*

by Hugo Lagercrantz and Theodore A. Slotkin

At first thought, being born would seem to be a terrible and dangerous ordeal. The human fetus is squeezed through the birth canal for several hours, during which the head sustains considerable pressure and the infant is intermittently deprived of oxygen (by the compression of the placenta and the umbilical cord during uterine contractions). Then the neonate is delivered from a warm, dark, sheltered environment into a cold, bright hospital room, where some large creature holds it upside down and in many cases slaps it on the buttocks. In addition, during the strains of birth—particularly hypoxia (oxygen deprivation) and pressure on the head—the fetus produces unusually high levels of the "stress" hormones adrenaline and noradrenaline—higher than in such severely taxed adults as a woman giving birth or a person having a heart attack. Adrenaline and noradrenaline, which are major representatives of a chemical class called catecholamines, typically prepare the body to fight or flee from a perceived threat to survival (a stress), and high concentrations of these hormones often indicate that the organism is in danger.

In spite of surface appearances, the stresses of a normal delivery are usually not harmful. Evidence collected by us and by others during the past two decades indicates that the fetus is well equipped to withstand stress, even early in gestation; indeed, it is the catecholamines that afford much of the protection from such adverse conditions as hypoxia. Moreover, it is actually important to undergo the events eliciting the production of stress hormones. The resulting surge of hormones prepares the infant to survive outside the womb. It clears the lungs and changes their physiological characteristics to promote normal breathing, mobilizes readily usable fuel to nourish cells, ensures that a rich supply of blood goes to the heart and brain and may even promote attachment between mother and child.

The first major clue that catecholamines are important to fetal survival came from pioneering studies done in the late 1960's by Robert S. Comline and Marian Silver of the University of Cambridge. They found that in the fetal cow, horse and sheep the adrenal glands (which sit above each kidney) produce catecholamines in response to hypoxia, even prior to the innervation of the glands by the sympathetic nerves. The finding was remarkable. Mature animals attain high catecholamine levels only when the sympathetic nervous system is activated and its splanchnic nerves, which arise in the central nervous system, stimulate the adrenal gland to secrete stress hormones into the blood. The sympathetic nervous system, together with the parasympathetic system, innervates all the major organs and helps to regulate such involuntary processes as breathing, circulation and temperature control.

When the central nervous system perceives a threat, it sends impulses to sympathetic nerves that release noradrenaline at target tissues, initiating rapid, localized "fight or flight" responses. In addition splanchnic nerves, which transmit signals by means of the neurotransmitter acetylcholine, activate the adrenal medulla (the inner part of the adrenal gland), causing it to secrete adrenaline and, to a lesser extent, noradrenaline; these hormones then travel through the blood to give rise to slower but more generalized and prolonged arousal. In particular the catecholamines typically accelerate the heart rate and output, increase the force of heart-muscle contraction and selectively dilate or constrict blood vessels to shunt blood away from organs that are nonessential during the threat (skin, intestines, kidneys) and toward ones that are essential (heart, brain, skeletal muscle). The stress hormones also dilate the bronchioles to aid respiration, cause fat and glycogen to be broken down into readily usable fuels, dilate the pupils and make the hair stand on end.

The Cambridge findings led to a search for the mechanism by which catecholamine levels become elevated in the fetus. The findings also raised a number of questions. What is the normal pattern of catecholamine production throughout gestation? By what processes do the catecholamines protect the fetus during oxygen deprivation? Do the effects of catecholamines differ between the fetus and the adult?

Fetal catecholamine production in the absence of a competent sympathetic nervous system was found to be a direct response (one not dependent on activation by nerve impulses) to stress by the adrenal medulla. The gland is proportionally larger than it is in the adult and produces more noradrenaline than adrenaline throughout most of gestation. Noradrenaline and adrenaline generally have similar effects, except that noradrenaline does not, for instance, cause blood to flow to skeletal muscle and does elicit a reflex slowing of the heart rate (mediated by the parasympathetic vagus nerve) instead of an increased heart rate. The fetus was also found to have extra sources of noradrenaline: tissues referred to collectively as paraganglia. These specialized tissues, which are red-brown nodules peppering the outside of the aorta, contain significant amounts of noradrenaline. They disappear during childhood.

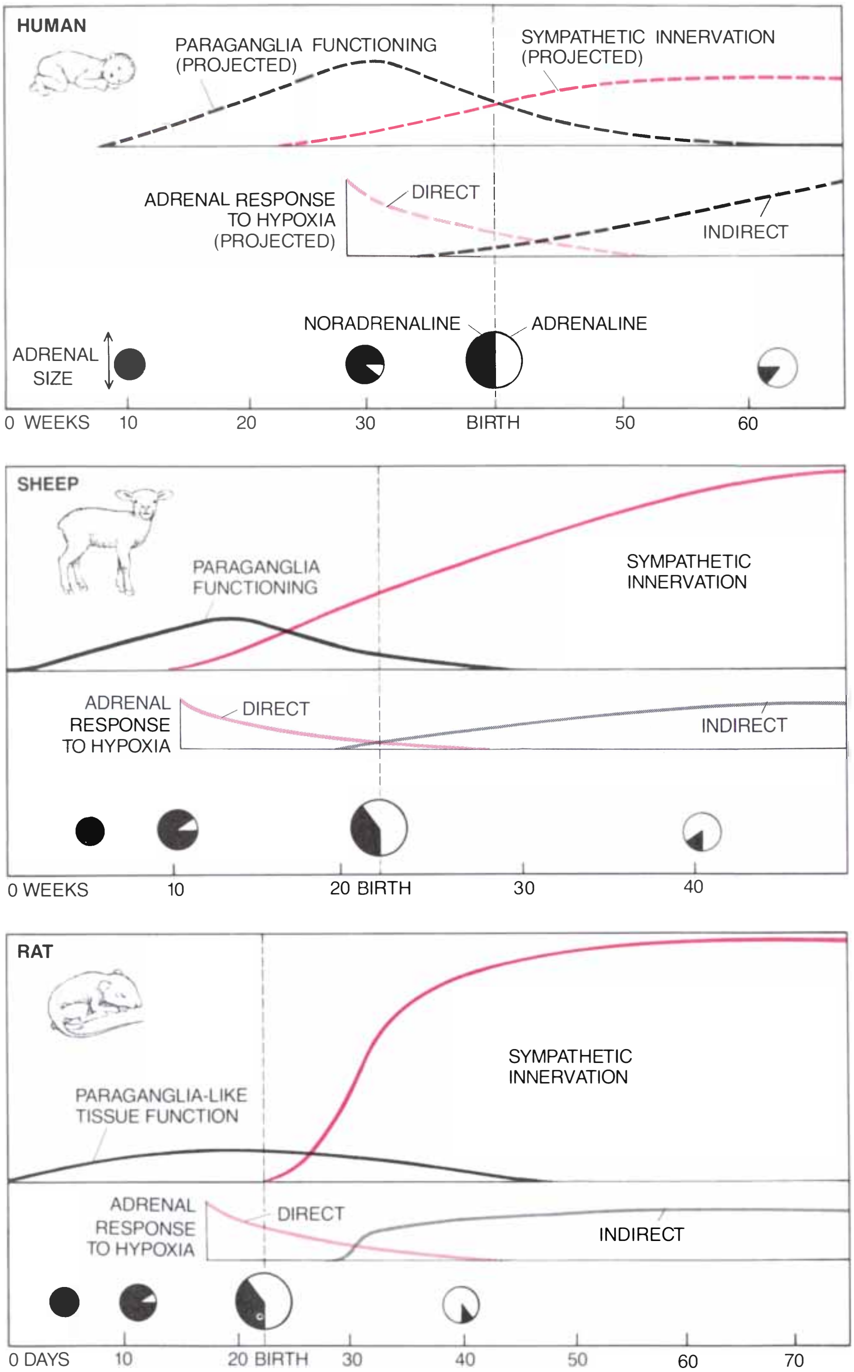
At the Duke University School of Medicine, Frederic J. Seidler, Jorge V. Bartolome and one of us (Slotkin) demonstrated that the development of the nervous system, particularly the nerves leading to the adrenal gland, somehow results in the adrenal gland's losing its ability to secrete catecholamines in direct response to stress. Our experimental subjects were newborn rats that develop a sympathetic nervous system only after birth and therefore rely exclusively on nonneuronal sources of catecholamines, particularly the adrenal medulla. We first showed that the animal with an immature nervous system is relatively well protected against hypoxia by catecholamines. Under low-oxygen conditions newborn rats secreted large amounts of catecholamines and were able to survive even intense and prolonged hypoxia. When the adrenals were removed at birth, or when we administered drugs that blocked the effects of catecholamines, the ability to survive hypoxia was lost.

The finding that the maturation of the nervous system itself leads to the adrenal gland's loss of ability to respond directly to stress was made when we forced the rats' sympathetic nerves to become competent prematurely. The manipulated rats thereupon were able to release catecholamines only by nerve-dependent adrenal secretion, and they were less resistant to hypoxic challenge than control rats. In addition, when adrenal innervation was delayed by surgically cutting the nerve supply before it became competent, the fetal direct-response mechanism—and its protective effect on survival—persisted.

In rats, where sympathetic innervation develops rapidly after birth, we can demonstrate that the adrenal medulla responds directly to stress until a few days after birth. In sheep, where innervation begins by the middle of gestation and is almost complete at birth, the ability of the adrenal gland to respond directly to hypoxia has virtually disappeared by birth. The newborn human is not as mature as the lamb but is more so than the rat, and the human's adrenal gland can probably respond directly to hypoxia at birth. The exact timing of full sympathetic-nerve maturation and the adrenal changeover is not known, however.

The natural pattern of catecholamine release throughout gestation was determined in two ways. One of us (Lagercrantz), working with colleagues at the Karolinska Institute in Stockholm, measured metabolites, or breakdown products, of adrenaline and noradrenaline in the amniotic fluid surrounding the human fetus. Resting

**NEWBORN WITH MOTHER is shown soon after a normal, vaginal birth. The vaginally delivered infant produces a high level of the so-called stress hormones adrenaline and noradrenaline, which belong to a chemical class known as catecholamines. Evidence indicates that this hormone surge during delivery enhances the ability of the neonate to survive once the umbilical cord is cut and that infants who do not experience a surge may be at a disadvantage. The surge also protects the infant from asphyxia during delivery.**




**DEVELOPMENT OF CATECHOLAMINE SOURCES is plotted in the rat, sheep and human fetus. In the mature organism (not shown) the main sources of plasma catecholamines are noradrenaline-secreting sympathetic nerves and (to a lesser extent) the adrenal glands, which produce more adrenaline than noradrenaline. The adult adrenal glands are controlled by the splanchnic nerves, sympathetic fibers that transmit signals by means of acetylcholine. In the fetus the sympathetic nerves do not function early in gestation. Until they do, the developing infant relies primarily on the adrenal glands for catecholamines. The fetal adrenal glands (which are proportionally larger than they are in the adult and secrete mainly noradrenaline) are at first not dependent on activation by nerve impulses but release catecholamines directly in response to stress; later this direct response of the adrenal gland fades. The fetus also has extra noradrenaline-producing tissues called paraganglia, which are least prominent in the rat and most prominent in the human. In the rat (*bottom*), which develops sympathetic nervous function after birth, the adrenal gland is able to respond directly to stress until a few days after birth. In the sheep (*middle*), which begins to develop a sympathetic nervous system midway through gestation, the adrenal gland all but loses its direct-response ability by the time of birth. In the human (*top*) sympathetic innervation is believed to progress more slowly than it does in sheep, and the adrenal gland may retain its ability to respond directly to stress for some weeks after the infant's birth.**

levels of metabolites proved to be fairly low, but the levels increased toward the end of gestation and were clearly elevated during distress, as when maternal toxemia caused life-threatening retardation of fetal growth in utero.

Such indirect measurement is certainly helpful, but the definitive evaluation of catecholamine response in the fetus calls for direct monitoring. Colin Jones and his group at the University of Oxford therefore chose to investigate catecholamine release in an animal model. Leaving the placenta attached, they removed the sheep fetus from the ewe, introduced catheters and electrodes into the fetus and then reinserted it into the womb. Jones found that catecholamine concentrations in the resting sheep fetus were very low until the last few days before birth. The concentrations increased a hundredfold during asphyxia, with the hormones coming primarily from the adrenal medulla.

Jones, and others too, described the process by which catecholamines protect the fetus from asphyxia. When the sheep fetus was given an injection of noradrenaline in a concentration corresponding to that triggered by hypoxia, the injection caused blood to be shunted toward the heart, brain, adrenals and placenta and away from organs that were not vital during hypoxia. These hemodynamic changes led to the elevation of blood pressure and in turn to an immediate reflex slowing of the heart rate. (This slowing of the heart is mediated by the parasympathetic vagus nerve, which becomes functional earlier than the sympathetic nerves do.) The reduction in heart rate promotes survival by reducing the heart's work and its need for oxygen. K. G. Rosén and Ingemar Kjellmer at the University of Göteborg and our group at Duke have further shown that interference with the cardiovascular effects of catecholamines prevents these adaptational changes during hypoxia, leading to circulatory collapse and death. (The effects of catecholamines on the oxygen-deprived fetus mimic the diving reflex seen in marine animals. When a diving animal dips its face in water, the blood flow is immediately restricted to all but the most vital organs, and the resulting blood-pressure increase triggers a lowering of the heart rate.)

The stress response of the fetus is more limited than that of the adult. The adult, who can fight or run in response to stress, is equipped to increase the heart rate and shunt oxygen-rich blood to muscles in response to stress. The fetus, on the other hand, needs to withstand oxygen deprivation, and it is exquisitely equipped to do just that and little else. In fact, the fetus is better able to withstand asphyxia than an adult, who typically develops heart-rate irregularities after a few minutes.

How might catecholamines reduce the heart rate in the fetus but not in the adult? One possible explanation is that the adult, unlike the fetus, produces much adrenaline in response to stress, whereas the fetus produces predominantly noradrenaline; elevated adrenaline levels can overcome a parasympathetic slowing of the heart, whereas noradrenaline cannot. (By the time of birth, the infant may release enough adrenaline to cause variable increases or decreases in heart rate.) A second reason the fetus is better able than the adult to withstand asphyxia may be that, as shown by one of us (Slotkin), the immature heart is less responsive: higher adrenaline levels are needed to elicit heart-rate increases, and so a catecholamine surge is less likely to cause irregular heartbeats.

The finding that catecholamine levels are low throughout most of gestation except during challenge suggested that, as in the adult, a catecholamine surge in the fetus serves as a highly effective protection system in emergencies. Intimations that the hormones not only protect the fetus during the acute stress of asphyxia but also prepare the body to function outside the uterus emerged from investigations into the possible overuse of Caesarean delivery.

Several years ago the Caesarean delivery rate in the U.S. and in Europe increased from 5 percent to as much as 15 to 20 percent. The reason was that new monitoring techniques made it possible to detect subtle changes in the fetal heart rate during uterine contractions; when complex changes were found during contractions (as opposed to simple heart-rate increases or decreases), the infant was often judged to be suffering from life-threatening asphyxia and was delivered surgically. On delivery, however, more than 50 percent of the infants were found to have few clinical signs of asphyxia.

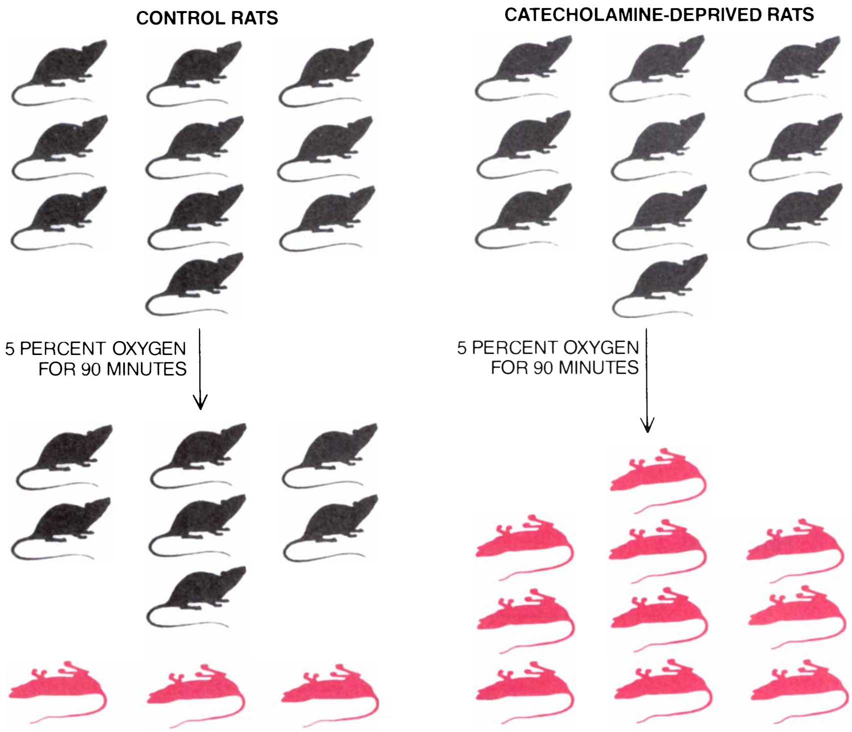
Concerned that many Caesarean sections might be unnecessary, one of us (Lagercrantz), with Peter Bistoletti and Lars Nylund at the Karolinska Institute, decided to find out whether a normal catecholamine release in response to maternal labor could account for the complex heartbeats in many fetuses. We found that catecholamine surges elicited by the normal birth process can indeed cause alterations in heart rate that might well be misinterpreted as signals of fetal distress unless the monitoring was supplemented by biochemical tests. Only when the *p*H level in blood from the fetal scalp fell below 7.25 did these apparent heart-rate abnormalities actually indicate asphyxia.

In exploring the monitoring issue we found that the normal birth process gives rise to a surprisingly large increase of plasma catecholamines in the human infant. Raul Artal of the University of Southern California has reported similar findings. Even at the beginning of normal delivery, when the mother's cervix was barely dilated (two to three centimeters), catecholamine concentrations in fetal scalp samples were about five times as high as the concentrations in a resting adult. After birth the catecholamine levels were found to have doubled or tripled again, indicating that they had surged during the next stage, when the mother was straining. (Catecholamine levels typically decline about 30 minutes after delivery, returning to resting levels about two hours after birth.)

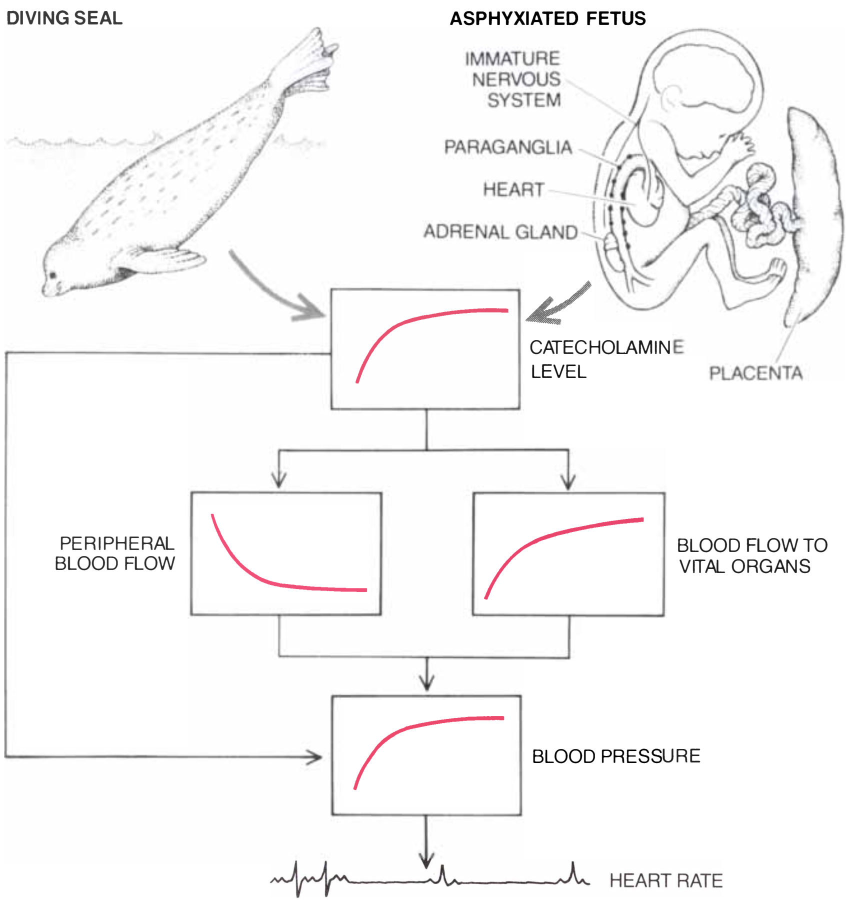
Studies further revealed that infants who were asphyxiated (*p*H below 7.25) because they were in the breech position or were being strangulated by the umbilical cord had catecholamine levels as high as between 500 and 1,000 nanomols per liter of plasma, which would cause a stroke in an adult. Such concentrations could only be achieved if asphyxia brought on further release of catecholamines from the adrenal medulla, paraganglia and nerves.

Part of the surge normally seen at delivery is likely due to some degree of hypoxia; nearly every newborn has an oxygen debt akin to that of a sprinter after a run. This oxygen deficiency, however, cannot alone account for the high catecholamine levels normally seen; a direct relation between asphyxia (as measured by plasma *p*H) and catecholamine levels is found only when an infant is considerably asphyxiated. Animal studies show that pressure on the head during uterine contractions can stimulate sympatho-adrenal activity, which suggests that compression of the head of the human fetus accounts in large part for the increased secretion of catecholamines during labor. Data collected by Lars Irestedt, Paul Hjemdahl and one of us (Lagercrantz) adds weight to the suspicion: infants delivered by elective Caesarean section without labor had low catecholamine levels, whereas those delivered by emergency surgery (after labor had begun) had a surge only slightly lower than that of vaginally delivered infants.

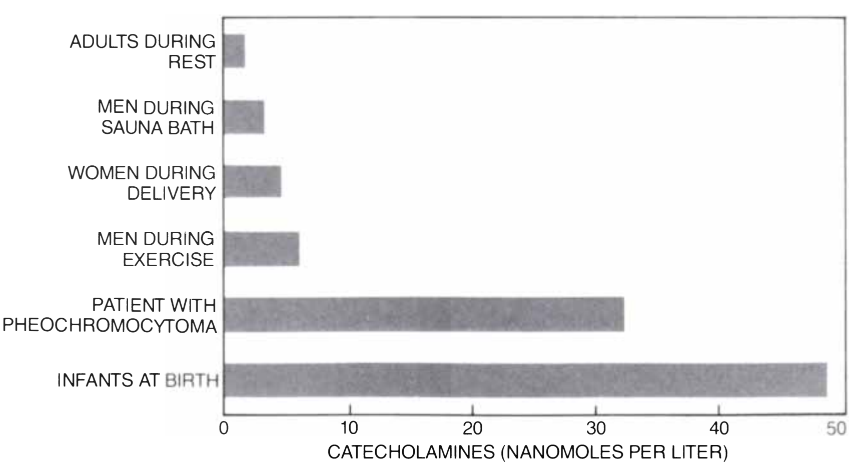
We could not ignore the surprisingly high catecholamine elevation observed in the fetus during even routine delivery. It implied that in addi-




**FIRST EVIDENCE showing that a catecholamine surge is required for the fetus or newborn to survive hypoxia (oxygen deprivation) is summarized. When day-old rats (which, like the human fetus, have an immature nervous system) had their adrenal glands removed or were given drugs that blocked the effect of catecholamines, none survived 90 minutes of a low-oxygen condition. In contrast, 70 percent of control rats survived the condition.**




**FETAL RESPONSE TO ASPHYXIA resembles that of a marine animal during a dive. When a seal submerges its head, the catecholamine level rises as the sympathetic nerves and adrenal glands secrete the hormones. The hormones then cause a decrease in peripheral blood flow and an increase in the amount of blood shunted to the brain, heart and adrenal glands. Together the catecholamines themselves and the blood-flow changes they produce lead to an increase in blood pressure and a slowing of the heart rate. Similarly, when a human fetus is deprived of oxygen, its adrenal glands and paraganglia (and, to a small extent, its developing sympathetic nerves) cause plasma catecholamines to soar. As a result the flow of blood to the periphery is decreased, blood flow to the brain, heart, adrenal glands and placenta is increased, blood pressure is elevated and the heart rate is slowed.**




**CATECHOLAMINE LEVELS in newborns and adults are compared. Levels in umbilical samples of newborns were found to be about 20 times higher on the average than levels in the venous blood of resting adults. Neonates also had higher catecholamine levels than a variety of stressed adults. (A pheochromocytoma is a catecholamine-producing tumor.) The finding that healthy infants who were delivered vaginally had higher catecholamine levels than severely stressed adults provided one of the first clues that a surge of hormones during delivery has a function beyond the acute protection of the fetus from oxygen deprivation.**

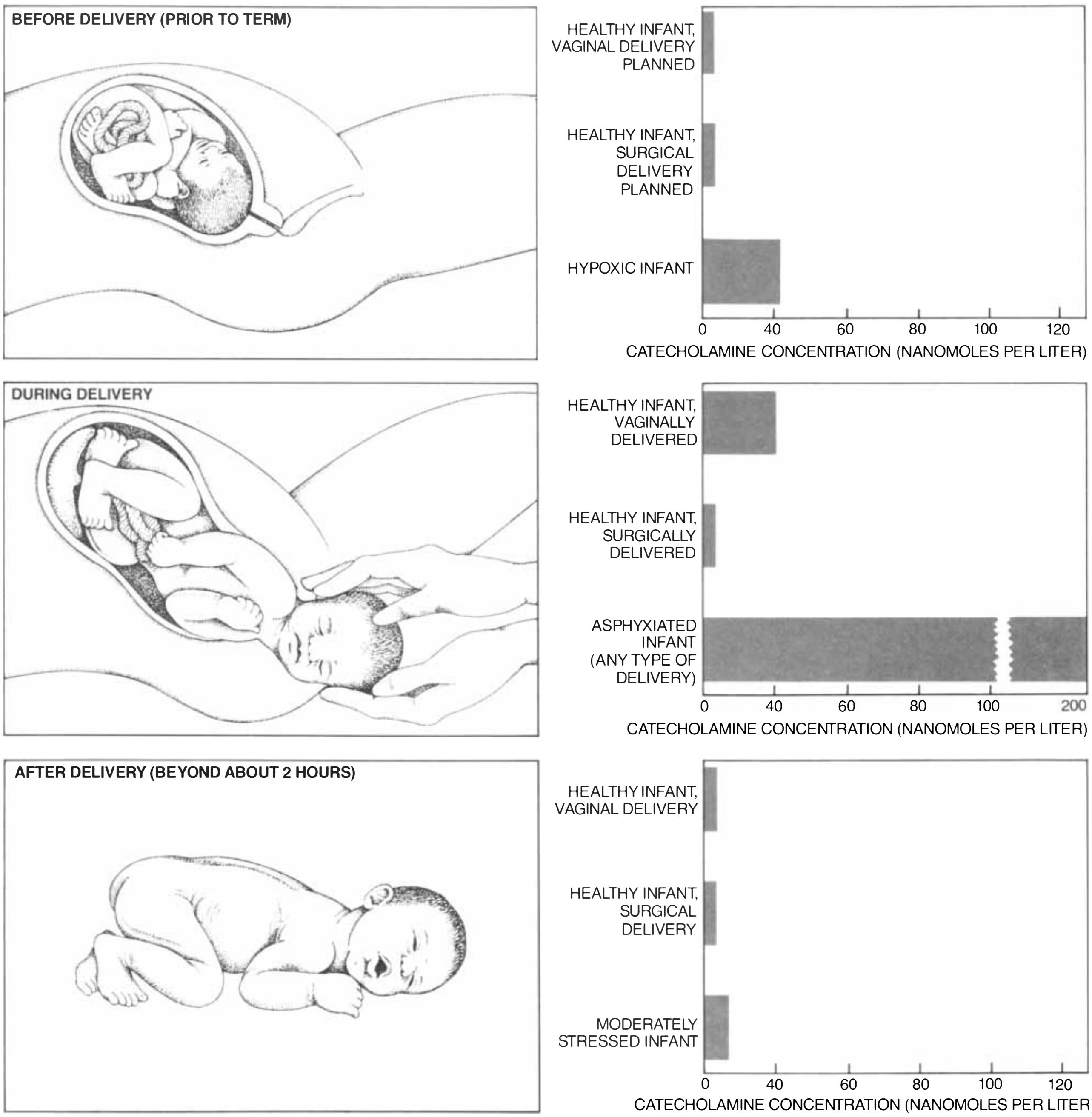
tion to protecting the fetus during the acute stress of delivery the catecholamine surge at birth probably has a second role: enhancing the neonate's ability to function effectively when it is first separated from the mother. Work in the past several years has demonstrated that the catecholamine surge facilitates normal breathing and has other effects that prepare the infant to survive lack of nourishment, oxygen deprivation or other adversity in the first hours of life.

The promotion of breathing is an extremely important adaptational effect of a catecholamine surge near delivery. Infants delivered by elective Caesarean section are known to be predisposed to breathing difficulties. Two major factors are inadequate absorption of lung liquid at birth, which gives the lungs a wet look, and inadequate production of surfactant: a soaplike substance that decreases surface tension in the alveoli of the lung (where oxygen and carbon dioxide are exchanged) and allows them to remain open. The absorption of lung liquid and the release of adequate surfactant both appear to depend on the sustained increase in plasma catecholamines in the hours immediately before birth.

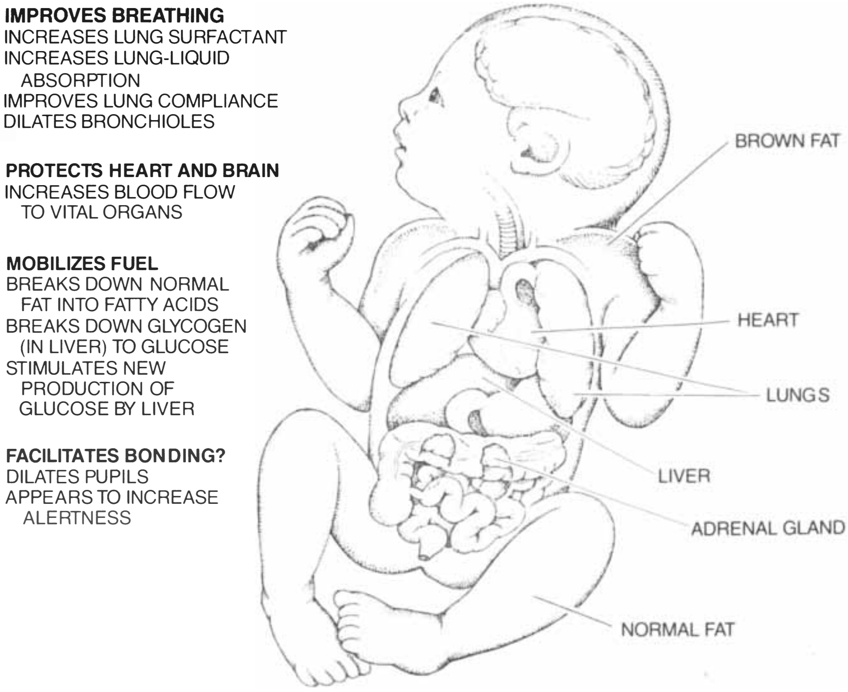
The importance of a sustained catecholamine surge (particularly adrenaline) to the absorption of lung liquid at birth has been demonstrated by elegant experiments carried out with fetal sheep by Richard E. Olver, Leonard B. Strang and D. V. Walters at University College London. Liquid is naturally secreted into the lungs of fetal sheep at the approximate rate of from 200 to 300 milliliters per 24 hours and moves through the lungs as if it were air. This phenomenon appears to be important for alveolar development during gestation but, as in the human, the process must reverse itself by the time of birth. When the workers injected the sheep with adrenaline, the secretion of lung liquid stopped immediately and the liquid was absorbed. The injected adrenaline had its greatest effect at term, as is to be expected if a catecholamine surge near delivery is important to the development of respiratory competence. Measurement of plasma adrenaline and lung-liquid absorption in fetal sheep during delivery confirmed that formation of liquid ceases and is replaced by absorption as adrenaline levels soar in the few hours before the infant is born.

The studies of rats at Duke confirm the vital importance of a catecholamine surge during delivery to lung function and the ability to withstand low-oxygen conditions during delivery and at birth. Administration of drugs that block beta-2 receptors, which mediate adrenaline's effects on surfactant production and lung-liquid absorption, severely compromised the ability of day-old rats to survive low-oxygen conditions. When slightly older animals (eight days old) were given such receptor-blocking drugs, no difficulties were observed. Apparently a catecholamine rise not only is crucial to lung function after birth but also must occur within a couple of days before or after delivery if it is to improve respiration in the neonate.

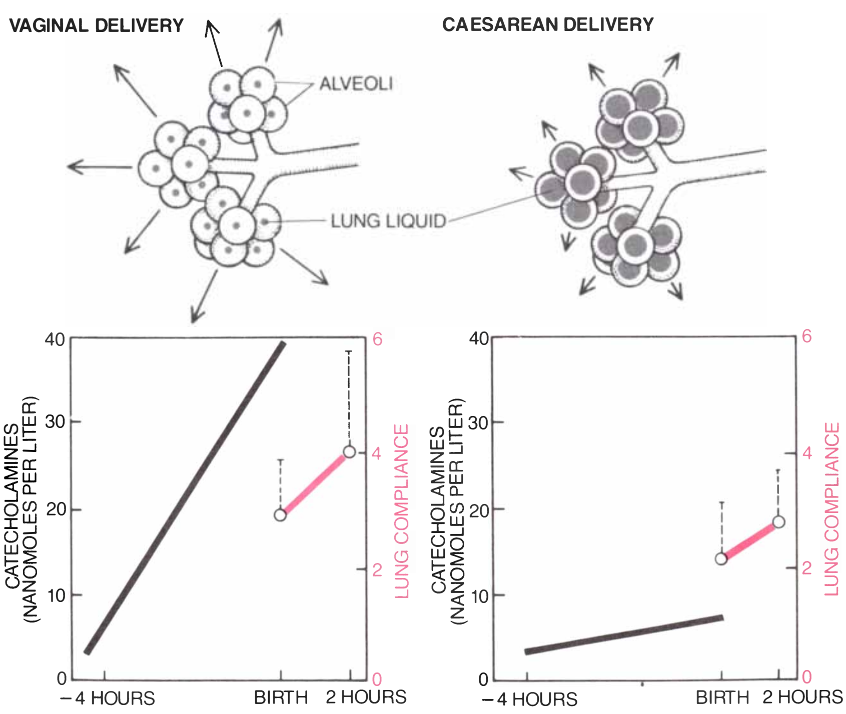
Although the production and absorption of lung liquid cannot be assessed directly in newborn humans, our Karolinska group (Irestedt, Gerd Faxelius, Kerstin Hagnevik, Bo Lundell, Ingrid Dahlin and Lagercrantz) has found that results derived from animal studies apply to human infants as well. We measured lung compliance (the ability of the organ to stretch and fill with air), which is partially dependent on lung-liquid absorption. There were few differences between vaginally and surgically delivered babies im-




**STRESS-HORMONE LEVELS in the human infant before, during and after delivery are estimated. (Direct measurement of catecholamines is not possible in the human fetus.) Levels are usually low until a few days before birth (*top*), rising only when the infant is stressed. They surge during a vaginal delivery (*middle*), largely in response to maternal uterine contractions (which cause intermittent oxygen deprivation) and to squeezing of the head as the fetus moves through the birth canal. This surge both protects the infant from asphyxia during delivery and facilitates stable functioning at birth. If a Caesarean section is done before the mother goes into labor, the infant is unlikely to have a catecholamine surge during delivery. If the infant is distressed during delivery, as it is when the umbilical cord wraps around the neck, catecholamine levels may soar extraordinarily high; such levels combined with low *p*H levels indicate that an emergency surgical delivery is needed. After the initial hours of life (*bottom*) catecholamines usually stay at resting levels.**




**ADAPTATIONAL EFFECTS of a catecholamine surge during delivery include promotion of normal breathing, alteration of blood flow to protect the heart and brain against potential asphyxia, immediate mobilization of fuel for energy and, possibly, enhancement of maternal-infant attachment. In general the effects prepare the body to sustain homeostasis (stable functioning) at birth even if the neonate is exposed to such adversity as oxygen deprivation. After delivery catecholamines can also activate, in response to cold, a specialized heat-producing tissue called brown fat that is unique to the infant and young child.**




**LUNG COMPLIANCE after vaginal and after Caesarean delivery is plotted, along with average catecholamine concentrations before birth, based on data from 27 infants. Compliance (the ability of the lung to stretch and efficiently exchange carbon dioxide and oxygen with capillaries through the alveoli) is measured by the relation of lung volume to pressure. Catecholamine concentrations before birth were measured in samples of scalp blood. Many studies have shown that infants delivered vaginally have more compliant lungs after birth.**

mediately at birth, but two hours later the vaginally delivered infants had significantly better lung compliance. Moreover, there was a close correlation between the catecholamine concentration at birth and lung compliance at the two-hour mark in the vaginally delivered group.

The fact that changes in lung compliance did not occur immediately after delivery but were delayed indicates that the effects are the result of circulating catecholamines; they are not caused by a mechanical squeezing of the thorax during delivery, which would have produced immediate effects. The conclusion that catecholamines facilitate breathing at birth is also confirmed by observation of infants born to mothers who have been given beta-adrenergic drugs, which mimic adrenaline's effects, in an effort to prevent preterm labor. Infants of such mothers rarely suffer from severe respiratory distress.

In addition to promoting normal breathing, a catecholamine surge before birth also speeds up the infant's metabolic rate at birth. This accelerates the breakdown of stored energy into forms that can nourish cells once the infant no longer receives a steady supply of nutrients from the umbilical cord. Mark A. Sperling of the University of Cincinnati College of Medicine has shown in sheep that formation of the sugar glucose begins immediately after a normal delivery. Similarly, our Karolinska group has found higher levels of glucose, free fatty acids and glycerol (all of them products of the breakdown of stored fuel) in the blood of vaginally delivered infants than in those delivered by elective Caesarean section. We have also found an increased incidence of low blood-sugar levels in surgically delivered babies, which indicates that the body's preferred source of fuel is in short supply in such infants.

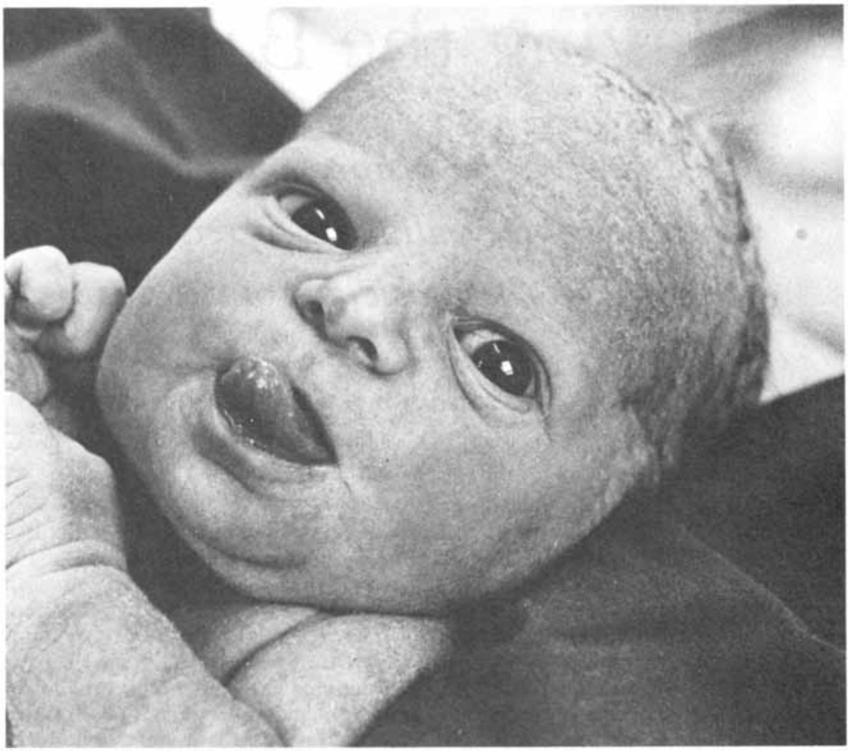
The third major adaptational effect of a sharp catecholamine rise during delivery is to alter blood flow. Our Karolinska group, in collaboration with Alice Yao of the Downstate Medical Center of the State University of New York in Brooklyn, has demonstrated that the vaginally delivered infant is born with an enhanced blood flow to vital organs and a restricted flow to the periphery. We found that the blood flow was lower through the leg and higher through the lungs in the vaginally delivered group during the first two hours of life and was correlated with plasma noradrenaline levels at birth. Being born with an increased blood flow to vital organs may not be critical to every neonate, but it should enhance the survival chances of a newborn who is experiencing breathing difficulties.

A review of Apgar scores, which measure the degree of asphyxiation in newborns, corroborates the importance of a catecholamine rise during delivery to the infant's ability to survive asphyxia. The review indicates that in general the higher the catecholamine surge is, the more the infant is likely to be able to withstand oxygen deprivation. Apgar scores, which are routinely assigned at birth to most infants born in North America and Europe, reflect the combined results of five measurements. The highest score, 10, is given if the heart rate is strong, breathing is regular, muscle tone is good, the infant reacts well to external stimulation and skin color is good. Infants whose score is below 7 are likely to have a lower heart rate, hypotonia (poor muscle tone) and cyanosis (blue coloring due to lack of oxygen in the blood), or some similar combination of ills.

At one time we believed there was an inverse relation between the catecholamine concentration in newborns and the Apgar score, since asphyxiated infants usually had higher catecholamine concentrations than healthy ones. Recent work has shown the relation to be more complicated. Infants who were moderately asphyxiated during delivery (according to *p*H levels in their umbilical blood) were found to have normal Apgar scores at birth if they had high catecholamine levels, low Apgar scores if they had low catecholamine levels. In other words, the secretion of catecholamines appeared to counteract the effects of oxygen deprivation.

More hypothetical than the finding that catecholamines enhance physiological adaptation to extrauterine life is the possibility that a catecholamine surge at birth facilitates attachment between mother and child. Marshall H. Klaus and John H. Kennel of the Case Western Reserve University School of Medicine have proposed that the first hour after birth is a critical period for maternal-infant bonding; if such attachment does not take place, the mother-child relation can be disturbed for some time. These findings are highly controversial. If they are correct, the implication is that being aroused and ready to interact at birth is more adaptive than being sluggish and unaware.

We suspect, based on admittedly informal observations, that the catecholamine surge at birth could give rise to such infant arousal. One reason for thinking so is that the catecholamine concentrations seen in a normally delivered neonate during the first hours of life would cause an adult to be alert and aroused and even at times to have a sense of well-being.



**INFANT IS ALERT five to six minutes after birth and has dilated pupils in spite of strong light, effects that appear to be caused by the surge of catecholamines during a vaginal delivery. By causing the infant to be alert, the hormonal surge may well facilitate attachment between mother and child during the first hour of life. This photograph is one of a series commissioned by John Lind of the Karolinska Institute for a study of infants at birth.**

Beyond their adaptational functions, catecholamines continue to be important to the newborn's survival during stress, just as they are for the fetus or the adult. If an infant is starved for several days (as it is in some cultures where the neonate is fed only after breast milk becomes available), a catecholamine response to the starvation will mobilize needed fuel from the liver's glycogen stores and from fat. If a newborn is left outside in the cold, a response by the sympathetic nerves leading to a unique tissue known as brown fat will stimulate the tissue to rapidly convert chemical bonds in its fat stores into heat. Brown fat, which is concentrated under the skin of the shoulder and near the kidney in the neonate, diminishes after early childhood. The neonate is particularly susceptible to rapid heat loss because it has a high surface-to-volume ratio.

Taken together, the weight of the evidence indicates that the elevation of "stress" hormones in the normally delivered newborn reflects not only a response to acute stress but also an attempt by the body to enhance the chances for survival at birth. Such findings suggest that infants delivered by elective Caesarean section before the mother begins labor may be at some disadvantage.

One way to overcome the disadvantage might be to administer catecholaminelike drugs to surgically delivered newborns, but this approach has yet to be proved safe and effective. In the meantime some obstetricians, particularly in the U.S., are now attempting to give the infant the benefit of a catecholamine surge by delaying surgical delivery—when they can—until the mother has experienced at least the early stages of labor. It is also evident that drugs interfering with the action of catecholamines (such as certain antihypertensives) should be avoided during pregnancy, and particularly near term.

It should be a comfort to parents to know that from the infant's standpoint the stress of labor during normal birth is likely to be less unhappy and more beneficial than common sense would suggest. Photographs commissioned for a study by the late John Lind of the Karolinska Institute bear out this point [*see illustration above*]. Lind found that most of the newborn infants in the series look aroused and cheerful—not unhappy—at birth.

# Playing the Baroque Trumpet

*Research into the history and physics of this largely forgotten instrument is revealing its secrets, enabling modern trumpeters to play it as the musicians of the 17th and 18th centuries did*

by Don Smithers, Klaus Wogram and John Bowsher

The current revival of baroque music puts great emphasis on methods of singing and playing that satisfy criteria for historically correct performance. In particular, many instrumental ensembles in Europe and North America are dedicated to playing 17th- and 18th-century music with authentic baroque techniques on original instruments or on modern instruments copied from surviving originals. The difficulties encountered in this attempt are nothing less than formidable. Considerable scientific and historical research is required—as much or more for the restoration of an antique piece of music as for the restoration of an old painting or building.

By now most baroque musical instruments have been successfully revived, the surviving originals having yielded virtually all their secrets. A few instruments, however, have been slow to attain the degree of revival needed for a pleasing and at the same time historically correct musical performance. Of these the baroque trumpet is the least understood. There appear to be two fundamental and not mutually exclusive reasons. The revival of an antique musical instrument is dependent on continuity in its design and manufacture and also on the persistence of its playing techniques in the memory of musicians. If there has been a significant period during which the instrument has not been manufactured, with an attendant loss of the skills necessary for playing it, there is little reason to suppose a revival will be either quick or easy. Today the baroque trumpet is not manufactured the way it was in the 17th and 18th centuries, and its playing techniques have been seriously compromised.

Unlike the modern trumpet, the baroque instrument had no mechanical devices (valves, keys, finger holes and the like) to help produce a complete scale from its lowest to its highest registers. It was in fact simply a long coiled or folded metal tube, rather like the modern bugle. And like the bugle, it had an air column whose overall length was fixed. Hence the instrument was more or less limited to those notes that can be produced from a fixed column of air; the notes are called harmonics or overtones. It is for this reason baroque trumpets and horns are sometimes referred to as natural instruments.

Restricted as he was to the natural tones—the harmonics—the baroque trumpeter had to have an instrument of appropriate length and design and had to develop special playing techniques to be able to sound enough of the available harmonics to generate a musical scale. Such an instrument would have been approximately eight feet long. (A shorter tube would yield fewer playable harmonics.) The eight-foot tube produces as its fundamental, or lowest note, C2 (two octaves below middle C on the piano). As a consequence the trumpet was capable of producing a discrete series of harmonic tones (called partials) above the fundamental note, which is also called the first harmonic [*see middle illustration on page 112*].

The fundamental can be altered by changing the length of the tube; the result is a higher or lower series of upper harmonics, depending on the pitch of the fundamental. The trombone is an example of a brass wind instrument with a variable fundamental. Because it has a slide that can vary the length of the air column, the trombone player can pick notes from several different harmonic series, thereby filling in the gaps of a single series to make a complete scale. In the modern trumpet and horn the valves serve a function similar to that of the trombone slide.

A problem for the baroque trumpet arises from the fact that several notes of the harmonic series are not in tune from the viewpoint of Western standards of intonation. This makes matters worse for the baroque trumpeter and constitutes the major obstacle in relearning the lost art of baroque-trumpet playing. The natural harmonics between the sixth and 16th partials that are prime numbers (evenly divisible only by themselves and 1) are notably out of tune for virtually any standard of musical temperament in the West. The seventh harmonic (and therefore also its octave) is too flat; the 11th is neither an F5 nor an F5 sharp, and the 13th is neither an A5 nor a G5 sharp.

Furthermore, such notes as D4, F4, A4, B4 natural, C5 sharp and E5 flat, which appear frequently in parts written for the instrument by 17th- and 18th-century composers, are nowhere to be found within a harmonic series based on a fundamental of C2. Yet the skilled baroque trumpeter was apparently expected to play them even though he had nothing but the standard, fixed-pitch instrument that under normal circumstances could produce only the series of harmonics described above. Until it was shown both in live performances and in recordings that these notes could in fact be played in tune and with enough resonance on an authentic baroque trumpet, many people thought they either were mistakes or were intended to be played on an instrument of variable pitch, such as a slide trumpet. Recent demonstrations on uncompromised historical instruments show that baroque composers knew what they were about.

The last writer who seems to have had some firsthand experience with the daunting problems of playing the baroque trumpet was the 18th-century German musician Johann Ernst Altenburg. He wrote that trumpeters needed both theoretical and practical instruction "partly because of these missing

**BAROQUE TRUMPETER** Johann Gottfried Reiche is seen here with his coiled *tromba da caccia*. Reiche was senior municipal musician and played first trumpet in J. S. Bach's Leipzig orchestra from 1723 until his death in 1734. The piece of music in his left hand is a short but difficult fanfare that has been recorded by one of the authors (Smithers) on a replica of the instrument; the recording is heard as the introduction to a CBS news program. The portrait, by Elias Gottlob Haussman, hangs in the old town hall in Leipzig.

and out-of-tune notes, but partly also because this instrument tends to remain a mystery, more so than other and contrived instruments." Recent scientific investigations—the first of their kind—have attempted to give some definition to the mysterious and to provide a body of data relevant to a better understanding of the technical difficulties of playing the fixed-pitch baroque trumpet the way it was intended to be played in countless works by Purcell, Bach, Handel and many other baroque composers. These investigations illustrate the inseparable union of science and practice.

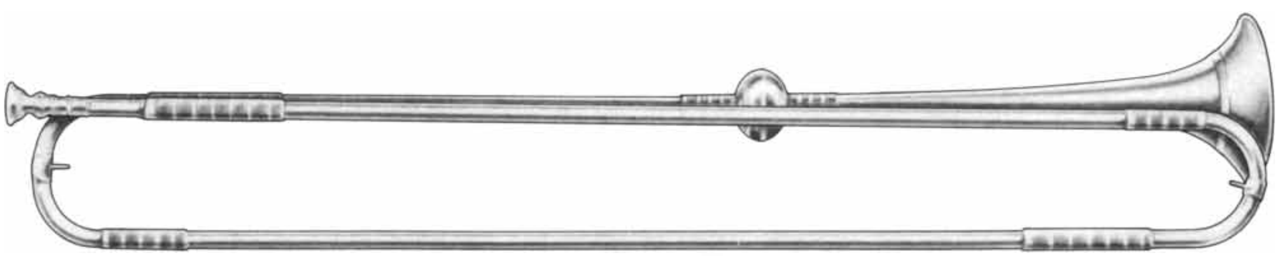
Earlier studies tended to concentrate on the acoustical characteristics of the instrument. Little attention was given to the integration of the instrument, its mouthpiece and the player. To understand the mysteries of baroque-trumpet playing requires an examination of the dynamic interactions among all three components, inasmuch as it is now known that the player has an enormous effect on the sound production of an instrument, considerably modifying its acoustical outputs. Such variables as an individual player's lips, teeth, tongue, oral cavity and throat also influence the sound directly, regardless of the particular physical characteristics of the instrument itself.

What occurs in the regions before (behind) and after (in front of) the vibrating lips is equally significant; the antevibratory and postvibratory environments are not mutually exclusive but interactive. In other words, the generation of a trumpet's sound is not entirely dependent on the region of vibrating air beginning in front of the player's lips and extending to the instrument's bell-flare termination. The vibrating lips set up regimes of oscillation in the air not only within the mouthpiece and instrument but also in the player's oral cavities. The way these resonances interact is now the subject of much study. The preliminary data make it clear that the interaction is a hitherto unexplained but important factor in the generation and control of certain notes, particularly in the middle and upper registers of the instrument.

For example, any skilled brass player knows that nimble manipulations of the tongue or the throat muscles, or both, on sustained notes make it possible to perform trills without recourse to any mechanical devices (valves or keys). Without necessarily understanding the underlying theory, the practiced artist thus alters the postvibratory resonances, which in turn control what is then heard from the player and his instrument. So-called lip trills are therefore a misnomer: fluctuations of the resonances behind the player's oscillating lips (in the antevibratory regions) are sufficient to alter the dynamics of the postvibratory air column (within the mouthpiece and the instrument). Comparable effects can be generated by displacements of the vibratory apparatus itself, either by shaking the mouthpiece on the lips or by some movement of the lower jaw. Such methods are generally unsatisfactory, however. They cannot attain the kind of control that can be exerted by subtle fluctuations of the antevibratory resonances within the player's mouth and windway.

In contrast to all other types of musical instrument, brass wind instruments have no oscillator of their own. The sound is produced entirely by the vibrations of the player's lips, which in turn modulate the air contained within the entire system. This is why the influence of the player on such acoustical parameters as sound and pitch is highly distinct with these instruments.

The integration of player, mouthpiece and brass wind instrument can be represented by an analogous but simplified model of these three distinct components in an electrical diagram [*see top illustration on page 112*]. Theory holds that for a trumpet player's lips to oscillate, the total of postvibratory and antevibratory impedances should have equal magnitudes but opposite angles of phase. With an instrument this condition is easy to meet, and



**TWO TYPES of baroque trumpet are depicted. At the top is a long, twice-folded natural trumpet (lacking valves or keys) tuned to a high-pitch D. The instrument is one of a matched set of three trumpets made by Johann Leonhard Ehe in Nuremberg in 1746. The other instrument is a circular *tromba da caccia* tuned to a low-pitch D. It was made by Johann Wilhelm Haas in Nuremberg in 1688. The Haas firm of trumpet, horn and trombone makers was reputed to have made the best natural trumpets during the baroque era. The trumpet (*tromba da caccia*) shown in the portrait of Reiche on the preceding page may be a Haas trumpet.**

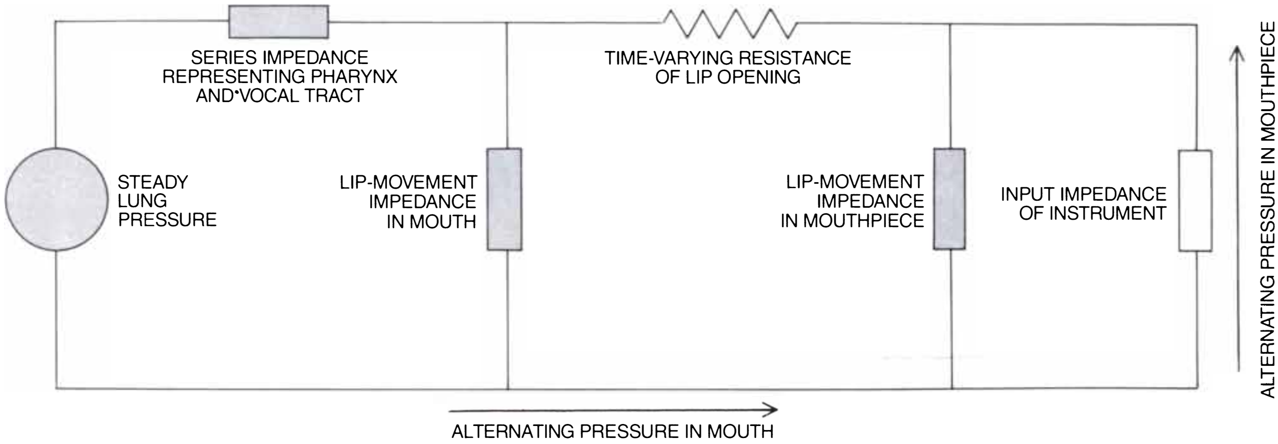
so the player can control the precise nature of the oscillation with small adjustments of the antevibratory impedances. When the player sounds his lips by themselves or with only a mouthpiece, however, the postvibratory impedances become those corresponding to free radiation or to the mouthpiece. The player in this situation has to make major adjustments of the impedances to ensure that oscillation can take place. Simple estimates of the magnitudes involved emphasize how important the impedances under the control of the player are and how relatively unimportant the impedance of the instrument itself is.

The least understood parameter of brass-wind-instrument playing in general and of baroque-trumpet playing in particular is the dynamics of the player's lips. Since the lips make the principal (and initial) contribution to the generation of sound in any brass wind instrument, it is ironic and regrettable that the dynamics of the lip vibrations is poorly understood. What one must deal with in this analysis is the nonlinear relation between the flow of air through an orifice and the pressure across the orifice.

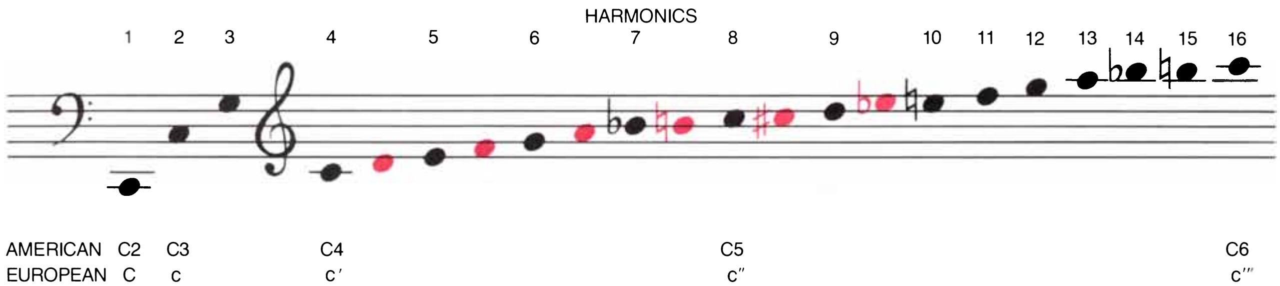
The orifice formed by the gap between the lips becomes smaller as the player produces higher notes. With the rise in frequency there is a disproportionately rapid rise of resistance, while at the same time the vibrating mass is reduced, with an attendant reduction of the surface area in the source of vibration and a decrease of amplitude.

A result of these phenomena is that there is an inherent difference between notes played with a large mouthpiece and notes played with a small one. This conclusion is entirely supported by the experience of playing a baroque trumpet in the original way. With a large, flat-rim mouthpiece it is possible to play notes in the high clarino register (from the 12th harmonic upward) that require different surfaces of the lips to vibrate than the lower notes of the *basso* register do. It is noteworthy that the baroque trumpet had about the same length of air column as the modern trombone has and a mouthpiece of comparable dimensions, and yet the clarino register of the instrument is at least an octave higher than the highest notes most trombonists today ever have to play.

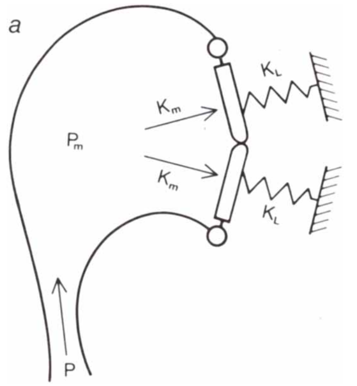
What the baroque trumpeter had to achieve in playing these high notes bears little resemblance to the techniques used by most modern trumpet players. Since modern mouthpieces are much smaller than those made in the 17th and early 18th centuries,

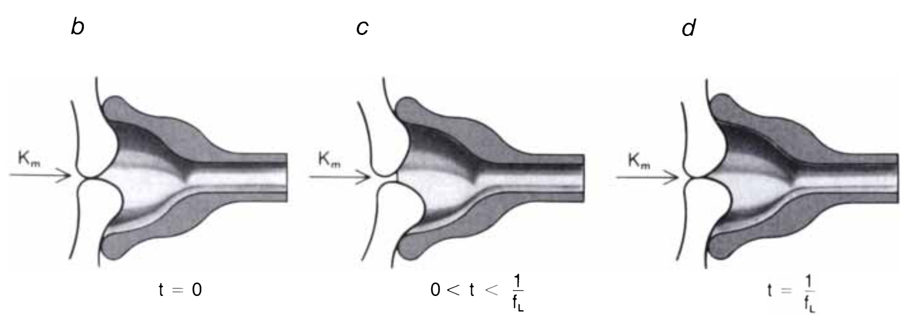


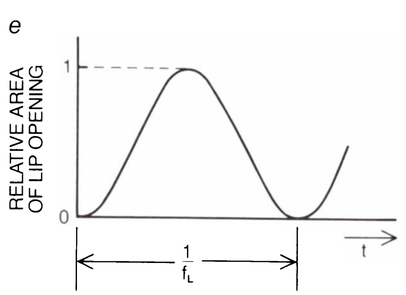

**CIRCUIT DIAGRAM** represents the integration of player, mouthpiece and brass wind instrument. Steady pressure comes from the player's lungs in series with the pharynx and vocal tract, represented by the series impedance. That impedance contains all the adjustments under the control of the player, such as jaw opening and tongue position. The player's lips move into and out of the mouthpiece, creating volume flows under the pressures in those two positions, shown as lip-movement impedances. The time-varying resistance of the lip opening reaches infinity when the lips are closed; the values are small when the lips open wide in playing low notes. This large cyclic variation results in the characteristic waveforms of pressure found in the mouthpiece. They are shown here as the mouthpiece pressure developed across the input impedance of the instrument and as the air pressure measured in the player's mouth.




**HARMONIC SERIES** based on a fundamental, or lowest note, of C2 (two octaves below middle C on the piano) indicates the range of notes available to a baroque trumpeter whose instrument had an eight-foot tube. The instrument could produce a discrete series (*black*) of partials, or harmonic tones, above the fundamental. Several notes (*color*) appear frequently in parts written for the baroque trumpet by 17th- and 18th-century composers but are not part of the harmonic series based on C2. The trumpeter in baroque times was apparently expected to play them regardless of the fixed pitch of his instrument. Recent research shows that they can be played.




**ACTION OF LIPS** in playing a brass wind instrument is represented in this schematic diagram, in which $P$ is the air pressure in the lungs, $P_m$ the air pressure in the player's mouth, $K_L$ the muscle force of the lips, $K_m$ the force of the compressed air in the mouth against the lips, $f_L$ the oscillation frequency of the lips and $t$ time. The lips represent an oscillating system composed of a mass $m$ and a spring (*a*). The lips behave like two flaps that tend to be held closed by the springlike elastic lip forces (*b*). As the internal air pressure in the mouth rises, the lips open at the moment the force $K_m$ becomes larger than the force $K_L$ (*c*). A certain amount of air then escapes as a positive pressure impulse. The resultant drop of pressure in the mouth causes the lips to close (*d*); the pressure rises once more, forcing the lips apart again. The result is a particular regime of oscillation in which the variation in the area of lip opening with time is sinusoidal (*e*). If lip mass, lip tension and air pressure are kept constant, the relaxation time will decrease with a reduction of volume in the mouth cavity. The player can alter the volume by changing the position of the tongue. He facilitates the playing of low notes by depressing the tongue and of high notes by arching it up and forward.

high notes are often produced only by pressing or squashing the lips as air is forced through the tightly pinched subcutaneous tissue. This not only causes the lips to tire quickly but also may cause the actual damage to lip tissue noted by some players.

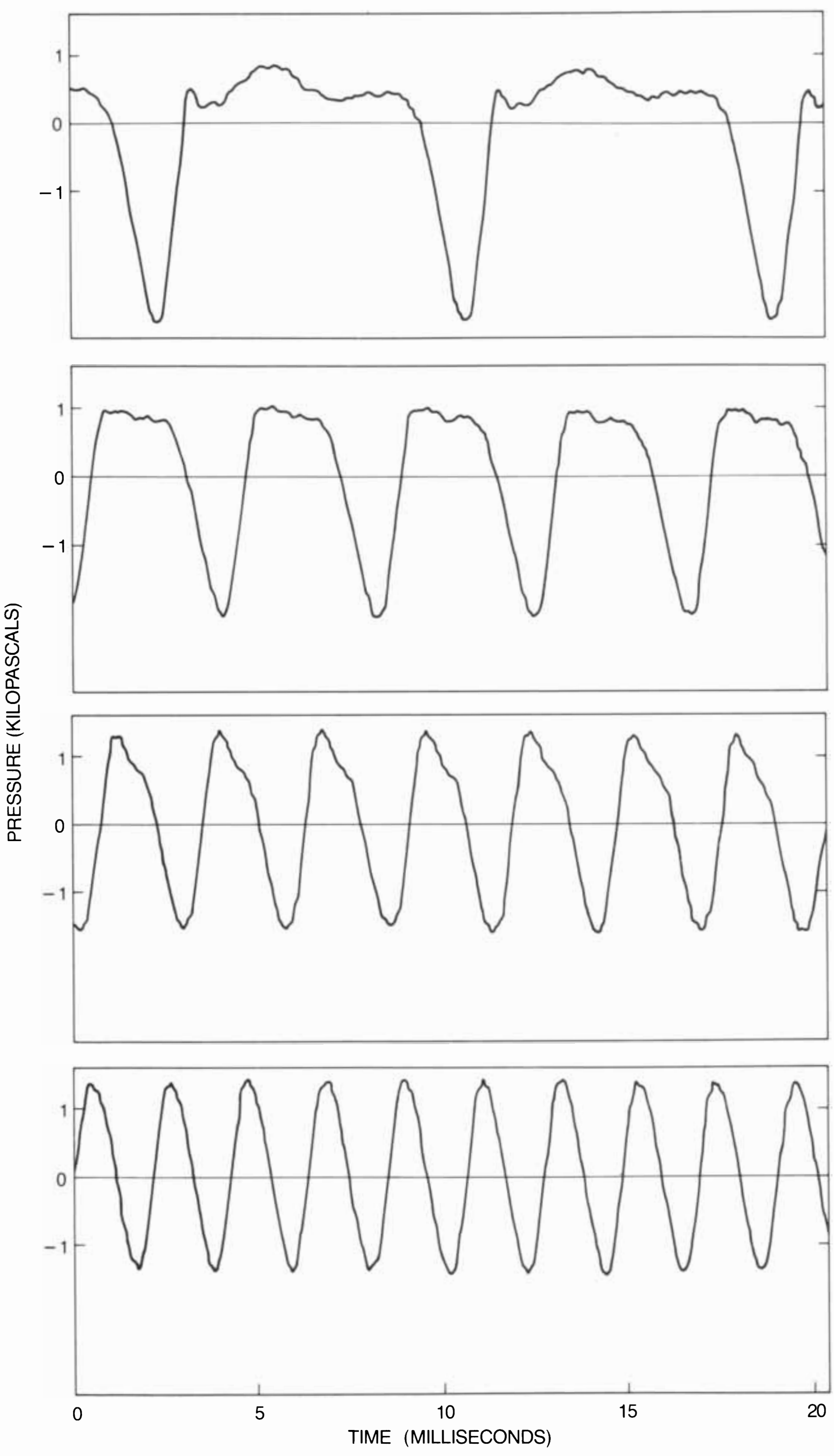
Because the lips are living tissue and the only source of sound for a brass wind instrument, there are strong correlations between playing the baroque trumpet and singing. The trumpeter's lips function much as the singer's larynx does, with the result that clarino playing has a number of similarities to the techniques used by a coloratura soprano. Indeed, many writings on playing a brass wind instrument assert that it is necessary for the student to study singing. One 18th-century writer's advice to the clarino player was to think always in terms of singing when playing and to imitate a beautiful voice as much as possible.

As many musicians who play horns, trombones or trumpets know, there are strong correlations between the first and third vocals (*aa* and *ee*) in singing and the production of notes from the lowest (*aa*) to the highest (*ee*) registers of a brass wind instrument. The baroque trumpet seems particularly susceptible to subtle changes in antevibratory resonances, whether because of the instrument's bore length of approximately eight feet or because of a combination of several acoustical factors, including those of the mouthpiece.

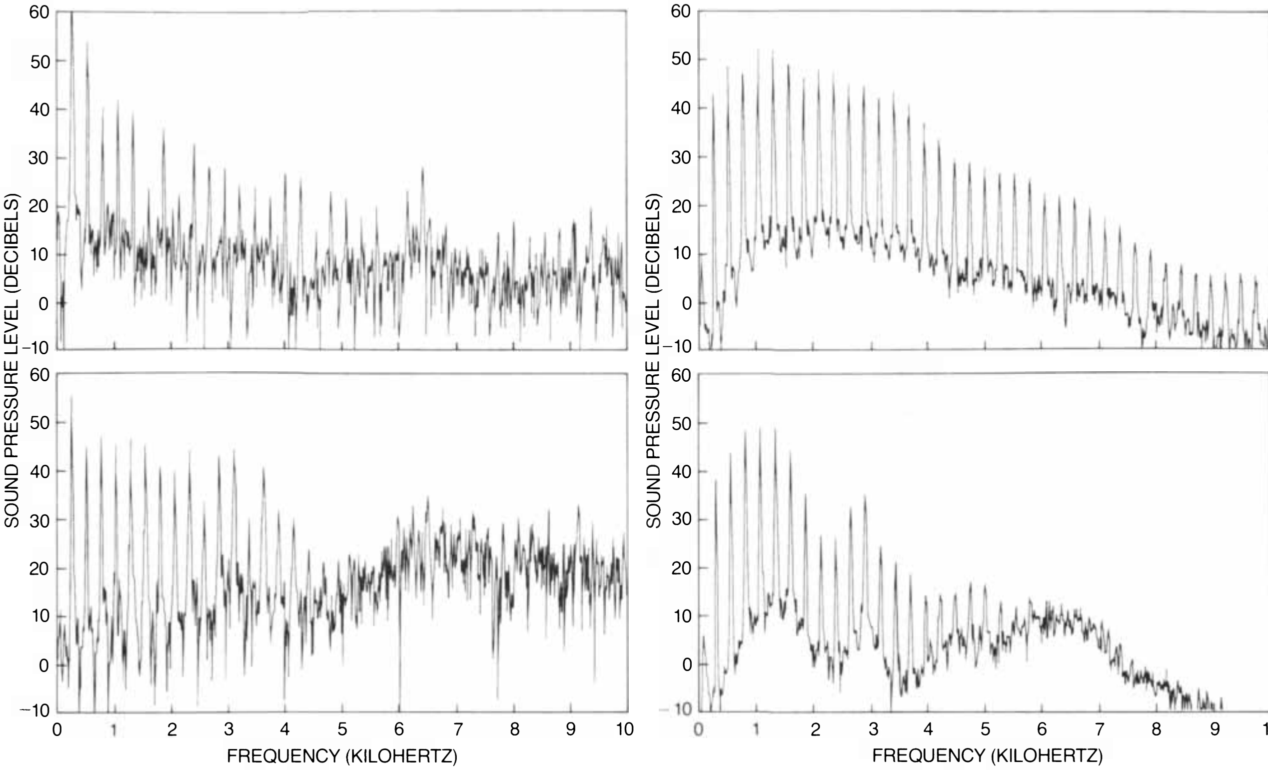
The production of notes in the lowest *basso* register depends on a wide-open windway, with the tongue depressed in the first vocal position and the lips as loose and massy as possible. High notes, on the other hand, require the tongue to be arched up against the hard palate (in the third vocal position), leaving as little space as is necessary to allow a flow of air and nearly closing off the windway from the oral cavity. In addition the lips must be tightly puckered.

Examination with a stethoscope of baroque-trumpet players performing from the lowest to the highest registers has verified the effect of these antevibratory resonances. Low notes were found to be sounded loudest in the laryngeal region and upper bronchi, whereas the highest frequencies attained their greatest amplitude in the uppermost parts of the throat and under the chin. Sounds also recorded at the cheeks and the bones near the nose may have been due both to air vibrations in the oral cavity and to the conduction of sound from the lips to the teeth and into the upper jaw.

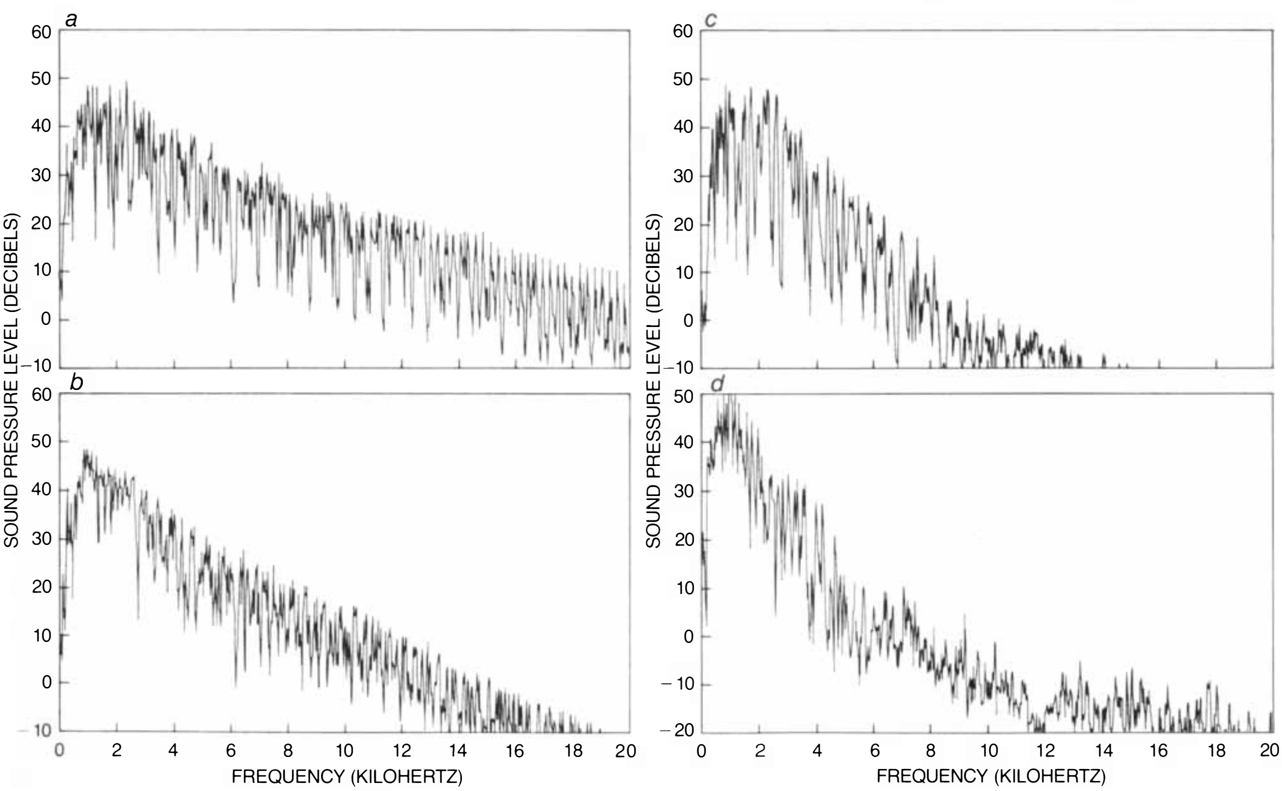
Changing the position of the tongue also affects the timbre, or tone color, of a note played on a brass wind instrument. The effect can be seen in recordings of the sound spectra for several comparable activities [*see top illustration on next page*]. When the tone C5 is recorded at a distance of one meter in front of a player sounding his lips without a mouthpiece or an instrument, the effect is virtually the same as when high and low notes are played on a jaws' harp. The number of overtones increases when the back of the



**ALTERNATING PRESSURE measured with a probe microphone in the mouthpiece of a trombone is charted for the notes B2 flat, B3 flat, F4 and B4 flat, reading from the top. The zero line on the pressure scale represents a pressure that is approximately atmospheric.**




**SOUND SPECTRA** for two methods of sounding the tone C4 are compared. At the left a baroque-trumpet player sounded his lips alone, without a mouthpiece or an instrument. At the right he sounded a twice-folded natural trumpet. In each case the top graph represents a large mouth volume (tongue back and down), the bottom graph a minimal mouth volume (tongue forward and arched).




**FOUR INSTRUMENTS** are represented in these sound spectra showing the average values of a two-octave tone array. The instruments are a twice-folded baroque trumpet in C (*a*), a coiled baroque trumpet, the *tromba da caccia,* or clarino, in D (*b*), a modern piccolo trumpet in B flat (*c*) and a cornetto (played with finger holes) in A (*d*). The differences between the baroque trumpets and the piccolo trumpet, which often substitutes in modern performances, show that baroque-trumpet music cannot have the necessary brilliance and clarity when it is played on the piccolo trumpet, which is one-fourth the length of the instrument the composer had in mind.

tongue is arched in the high, third vocal position (reducing the volume of resonance in the mouth), and it decreases when the tongue is depressed in the first vocal position (increasing the mouth's resonance volume). At the same time there is a rise in the noisy components, as is seen in the increase of 20 decibels in the sound level at frequencies above six kilohertz (6,000 cycles per second). This means that the resonance frequency of the mouth cavity rises with the elevation of the tongue and that there is a greater turbulence in the airflow and a decrease of lip displacement.

Comparable spectra appear when, instead of sounding the lips alone, the player blows a natural trumpet under identical conditions [*see bottom illustration on opposite page*]. For a large mouth volume (low tongue position) a full and bright sound is obtained; it has a rich assembly of overtones and only a slight amount of noise, because of the strong vibration and large displacement of the lips. When the mouth volume is made smaller with a high tongue position, the sound is rougher and has a nasal and obtrusive tone color. To obtain stable tones in the very high clarino register it is necessary to reduce the volume of the mouth as much as possible, although this action does not always produce a satisfactory tone color.

Next to the instrument itself, the mouthpiece is the most crucial element for playing a brass wind instrument. It is an integral part of the instrument, equal to it in importance for generating specific impedance maxima (the ratio of the pressure to the volume velocity at a given surface) and achieving a favorable intonation curve (the curve that reflects the values of the playing frequencies), particularly for notes above the sixth harmonic. The mouthpiece and the instrument therefore have to be matched. Moreover, as we have pointed out, the player is part of the total system, and so the mouthpiece must be matched to the player as well as to the instrument. It is the interface between the two.

For these reasons the mouthpiece is one of the player's most personal items of equipment. The relation seems to be more critical for playing the baroque trumpet than it is for playing any other brass wind instrument because of the extremes of range with which the baroque trumpeter has to cope. The instrument's lowest *basso* register is equal to that of the trombone. Its uppermost clarino register is often above the normal range of a modern trumpet, and that range is already an octave higher than the upper limit of the trombone and at least a fifth above the French horn.

The many ancient baroque-trumpet mouthpieces that survive suggest players and makers took a strong interest in mouthpiece design and construction. No two are alike. Mouthpieces were clearly made to suit the feel and facial structure of individual players. Equally important, mouthpieces would have been made to match the player's instrument.

A modern player can learn to cope with mouthpieces of different kinds, usually one at a time over a long period. The process carries some risk, however, to the sensitive neuromuscular apparatus of the sound generator (the lips as well as the neck and facial muscles). Given a choice, a competent performer always prefers to have a mouthpiece that feels comfortable and produces the best acoustical response.

Notwithstanding the uniqueness of each baroque-trumpet mouthpiece, the mouthpieces collectively have several common features. Nearly all of them are larger than the largest modern trumpet mouthpieces. Various methods of dating indicate that the ones from the 17th century are larger than ones made in the 18th century. More important, the design of baroque-trumpet mouthpieces is consistently different from that of their modern counterparts. The design not only promotes greater resonance in the lowest harmonics but also gives greater definition to the uppermost partials and allows more control of nonharmonic tones.

Furthermore, the peculiar design of some baroque-trumpet mouthpieces causes an increased excitation of the higher frequencies. That, in conjunction with the deeper and more resonant presence of the lowest harmonics generated in an eight-foot air column, produces for each note a far richer and broader sound spectrum than one hears from shorter modern trumpets. The modern piccolo valve trumpet, which is the instrument employed in most contemporary performances of baroque music, produces somewhat shrill tones because it is only about two feet long.

The last part of the equation of baroque-trumpet playing, but by no means the least important, is the instrument itself. Here one encounters an oddity: in spite of the perfection of modern, machine-made components as opposed to the irregularities in 18th-century handmade construction, the old instruments are much easier to play and are more in tune than modern facsimiles. Recent acoustical research has shown that the intonation of an instrument (the values of the resonance frequencies) is determined almost exclusively by the shape of the inner walls of the air column. On the other hand, the response of an instrument is almost entirely determined by what is called the $Q$ factor of the resonance, that is, by the smoothness and the degree of precise continuity of the inner surface of the metal tube (and to some extent by the material the tube is made of).

Baroque trumpets, both original and modern, consist mainly of joined cylindrical tubes (either coiled or folded), with a final short conical flaring section. In modern reproductions the methods of manufacture produce a consistently regular and smooth inner surface, thereby creating resonances with high $Q$ factors. Inasmuch as there is no particular relation between resonance frequencies and the harmonic components of the sound, the modern instruments display less variability of pitch than the 17th- and 18th-century originals. As a result it is difficult to play them in tune.

The components of old brass instruments were fashioned from sheets of brass, as they are today, but hand-hammered sheets of brass lack the consistency of modern ones rolled by machine. Furthermore, when the hand craftsmen made the sheets into lengths of tubing by shaping and hammering them over steel rods, they introduced many irregularities on the inner surface. These small but significant imperfections were compounded by irregular tube diameters, imperfectly made fittings and a lack of symmetry in the places where the tubing had to be bent. All these variations caused a decrease in the $Q$ factor of the baroque trumpet's resonances, thus flattening the resonance curve.

The end result is that the musician playing an antique baroque trumpet can "bend," or vary, the natural harmonics, so much so that many of the old instruments can be played in tune without creating unacceptable changes in tone color and response. To achieve this capability nearly all modern makers of baroque trumpets have resorted to the expedient of putting pitch-correcting finger holes in their instruments so that the trumpeter can play them in tune. This is not only a falsification of historical principles but also a self-defeating compromise of specific acoustical parameters and of the intended playing techniques. Only by consistently applying historical principles to all three parts of the equation—the player, the mouthpiece and the instrument—can one accurately revive the lost art of playing the baroque trumpet.

# Fiber optics: beyond 20,000,000,000 bits/sec.


## Summary:

**GTE scientists have demonstrated laser modulation and detection at rates as high as 20 gigabits/second. They have developed ultra-small lasers that have light emitting areas as small as 0.2 square microns. They are working to improve the glass fiber itself, as well as to produce optical analogs of electronic switches with the long-range goal of all-optical systems, in which message streams are switched as much as 10,000 times faster than at present.**


GTE commercial involvement in fiber optics communications systems dates from the first such installation in Artesia, California, in 1977.

Our scientists developed the system's technology and equipment, and have been contributing to the state of the art ever since.

Current projects deal with increasing the capacity, the versatility, the applications of the systems; longer-term, we are exploring the possibility of all-optical systems.

## Faster and faster...

Until recently, optical systems processed digital streams at speeds ranging up to hundreds of megabits per second.

Fast though that may seem, today's carriers are seeking speeds in the gigabits-per-second range. This might even permit the glass to be brought directly to satellite earth stations or microwave towers, for example, for direct conversion of radio signals to light.

Recently, GTE demonstrated the ability to turn diode lasers on and off at rates as high as 20 gigabits per second—about 333% higher than the greatest previously recorded speed.

## ...and smaller and smaller.

Such speeds require very special lasers. And, as you can see from the electron micrograph at upper right (the head of an ant looking at one of these lasers), it is extremely small.

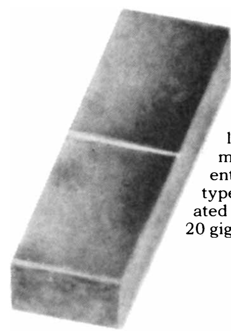
It was constructed on a wafer of InP, by epitaxial growth of a layer of InGaAsP approximately 0.1 micron thick. This was then etched to a mesa shape, and further layers of InP added.

The resulting laser cavity is approximately 0.2 square micron in area, and provides an excellent mate for single-mode glass fiber (fiber with a core of such small diameter that light travels a single path—mode—drastically lowering its dispersion within the fiber).

## Switching light with light.

In another project, we are investigating the possibility of ultimately eliminating the electronics altogether by using optical switches.

We are working with materials whose indices of refraction vary with the intensity of incident light—a non-linear response.



Head of an ant dwarfs a sub-micron-sized diode laser in this electron micrograph. GTE scientists developed this type laser, and have operated it at rates as high as 20 gigabits a second.

Ultimately, if it actually does become possible to switch systems optically, an improvement in speed of as much as 1,000,000% is theoretically possible.

In its brief history, fiber optics has made astonishing strides. At GTE, we are working to continue at the frontiers of this science—to make fiber optics an even more helpful technique to meet the endless needs of tomorrow's telecommunications.

The box lists some of the pertinent papers GTE people have published on various aspects of fiber optics. For any of these, you are invited to write GTE Marketing Services Center, Department FO, 70 Empire Drive, West Seneca, NY 14224. Or call 1-800-833-4000.

# THE AMATEUR SCIENTIST

## *Wonders with the retroreflector, a mirror that removes distortion from a light beam*

by Jearl Walker

An ordinary mirror will reflect a beam of light to its source only if the mirror is perpendicular to the beam. A retroreflector, on the other hand, returns a beam to the source regardless of the angle at which the beam strikes the device. With an array of small retroreflectors one can do a novel experiment: if the beam is distorted by, say, a ruffled sheet of plastic before it reaches the array, the array removes the distortion by reversing the beam back through the plastic. David M. Pepper of the Hughes Research Laboratory has written a manuscript, on which my discussion is based, describing this experiment and others that demonstrate the phenomenon. The experiments were based primarily on research by Harrison H. Barrett and Stephen F. Jacobs of the University of Arizona.

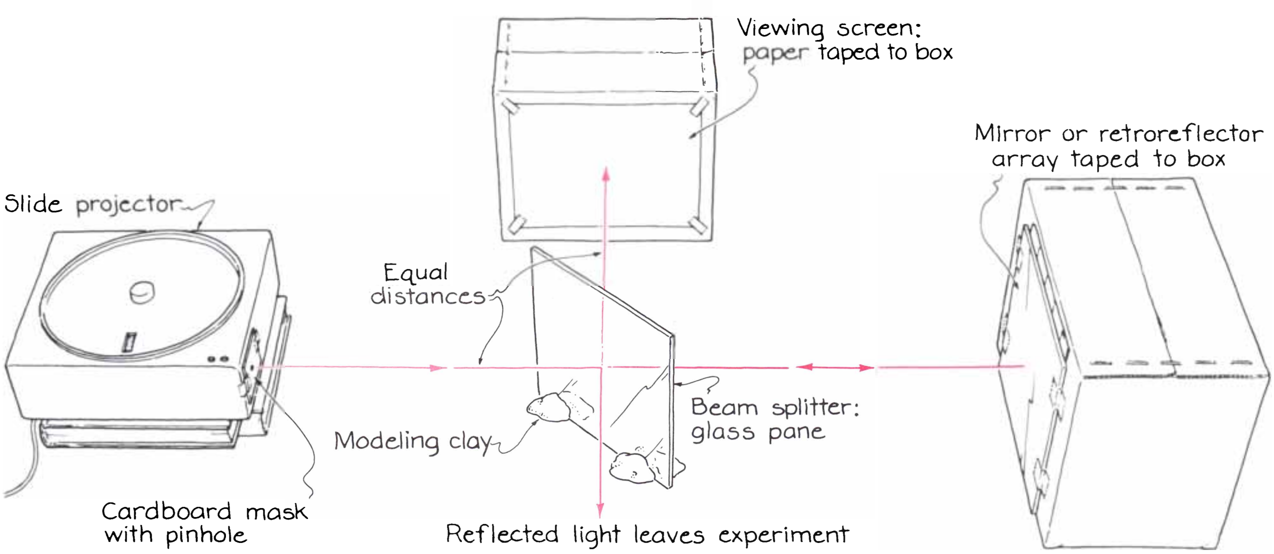
Retroreflecting arrays are available at sporting-goods stores and automobile-accessory shops in the form of flexible plastic sheets sold as safety reflectors. Sewn onto clothing, fastened to a bicycle, glued to a highway stripe or attached to a variety of other objects, the reflectors make the objects more visible at night. For example, when the headlights of an automobile strike an array, the light comes strongly back as a warning to the driver. The success of the application is actually due to an imperfection in the arrays. If the retroreflectors returned the beam of light precisely instead of spreading it slightly, the entire beam would return to the headlights rather than to the driver.

To do Pepper's experiments you should buy a retroreflecting flexible plastic sheet that is at least three centimeters square. If you can find only narrow strips, cut them up to form the square. Tape it onto a sturdy box. Set up a slide projector so that it shines on the array. If possible, remove the projector's front lens. Mount on the front of the projector a thick piece of cardboard through which you have made a hole two or three millimeters in diameter. The pinhole serves as a point source of light. If you do not have a projector, set up a bright flashlight or even an ordinary incandescent bulb to shine through the pinhole.

In order to see the shape of the returned beam, you must split off part of it and direct it to a viewing screen. You can do this by placing a glass pane in the beam at an angle of about 45 degrees. The glass should be clean and clear, but it does not have to be of excellent optical quality. Glass from a picture frame will do. Hold it in place with modeling clay or with the frame from which the cardboard backing has been removed.

When the beam from the pinhole reaches the glass, it is split into two beams. One beam continues on to the array and the other is reflected out into the room. The beam reflected from the array comes back to the glass, which makes another split. One beam passes through the glass, returning to the pinhole. The other beam is reflected from the glass to a sheet of paper that serves as a viewing screen. Since light is reflected from both the front and the back surface of the glass, the screen displays two slightly displaced images. You will get less displacement and a clearer image with a thin piece of glass serving as the beam splitter.

The beam splitter should be equidistant from the screen and the pinhole so that light returned by the array travels the same distance to each. When everything is arranged, you will see a small, bright circle on the screen. It is an image of the pinhole produced




*An arrangement for experiments with retroreflection* by light that would have returned to the pinhole except for the action of the beam splitter. The image is about the same size as the pinhole but has a fuzzy edge because of the double reflection by the glass and the imperfection of the retroreflecting array.

If you look at the screen through the beam splitter, you will see a bright spot that seems to be on the screen. You might well mistake it for the image returned by the array. Instead it is an image of the pinhole formed by your eye from the light reflected toward you by the beam splitter. Such images, called virtual images, are commonly produced by mirrors. In the experiment the beam splitter acts as a weak or semitransparent mirror. If you take the proper angle of view, you can align the virtual image with the real image on the screen. To avoid the complication of the virtual image do not look through the beam splitter.

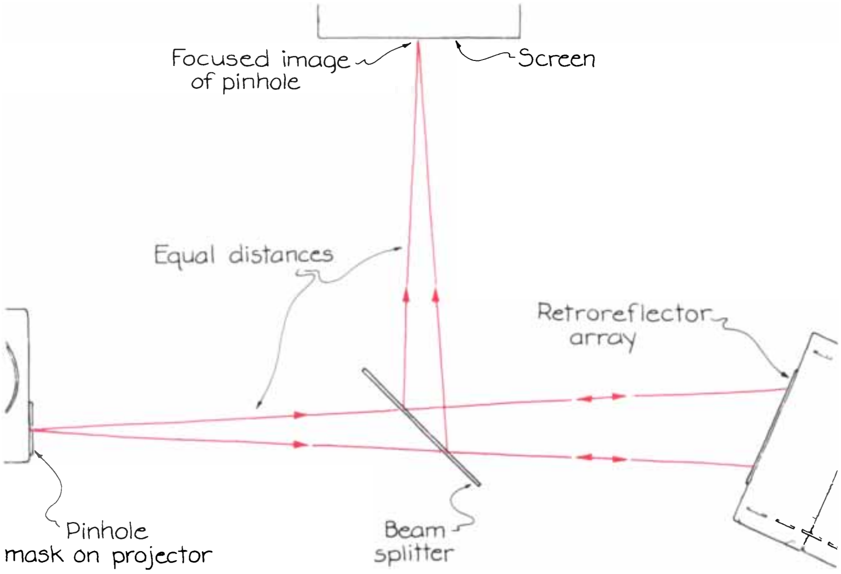
If you change the distance between the retroreflecting array and the beam splitter, the image of the pinhole on the screen does not change. Neither does it change if you slide the array across the beam. If you rotate the array with respect to the beam, the image does not change until the angle is large. In short, the array sends a beam of light back toward the light source even when the plane of the array is not perpendicular to the beam.

Replace the array with a flat mirror. The entire screen is bathed with dim illumination, and you will not see an image of the pinhole. Try a concave mirror. It produces a clear image of the pinhole on the screen only when its focus is on the pinhole. If you move the mirror along the beam or rotate it with respect to the beam, the image broadens, becomes distorted and finally disappears.

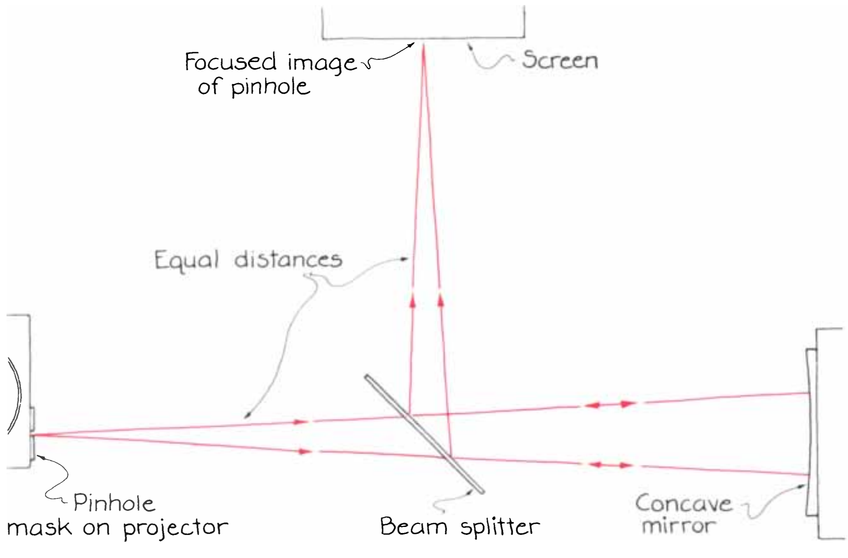
The action of these mirrors and of the retroreflecting array can be understood with the help of the illustration on the next page. Part *a* depicts the reflection by a flat mirror. The pinhole is represented by a point source of light. Waves of light flow from the point source. Each one displays a wavefront that is part of a sphere. One such wavefront is shown in cross section as part of a circle centered on the point source. As the wavefront travels away from the point source, the radius of the circle increases. When the wavefront reaches the mirror, sections of it are reflected to form a new wavefront traveling in the opposite direction. The reflected wavefront is part of a circle centered on an imaginary point source of light to the right of the mirror. The imaginary source is as distant from the mirror as the true source is. The reflected wavefront cannot form an image of the pinhole because it continuously grows larger, never attaining a focus. When the beam splitter directs part of the light to the screen, the entire screen is illuminated by the expanding wavefront.

Part *b* of the illustration demonstrates the action of the concave mirror. The mirror reflects and focuses the incident wavefront, but the light returns to the source only if the pinhole lies on the focal point of the mirror. Since the screen is as far from the beam splitter as the pinhole is, a clear image of the pinhole is formed on the screen. For other positions of a concave mirror the light spreads more broadly over the screen. If the mirror is not oriented symmetrically on the beam, the focusing of the light is skewed. The result is either a distorted image or no image.

Part *c* of the illustration represents the action of a highly specialized mirror system that is related to the commercial retroreflecting array. It is called a phase-conjugate mirror [see "Optical Phase Conjugation," by Vladimir V. Shkunov and Boris Ya. Zel'dovich, *Scientific American,* December, 1985, and "Applications of Optical Phase Conjugation," by David M. Pepper, *Scientific American,* Jan-




*A view from overhead of the focusing in a retroreflecting array*




*Focusing by a concave mirror*

uary]. This ideal retroreflector exactly reverses the light rays incident on it. Moreover, the reflected wavefront preserves the shape and orientation of the incident wavefront. With such an ideal mirror all the returned light that passes through the beam splitter travels through the pinhole. The fraction of the light that travels to the screen forms an exact image of the pinhole.

The retroreflecting sheets Pepper employs are less than ideal. Each retroreflector in a sheet returns a section of the wavefront, preserving its shape and orientation as is shown in part *d* of the illustration. Because the retroreflectors lie in a plane, however, the sections do not fit together to re-form the initial wavefront. Instead they are spread along a curved line that is identical with the wavefront reflected by the flat mirror. For this reason the sheets are said to be pseudoconjugators. Each section of the wavefront returns to the pinhole or to the image of the pinhole on the screen, but together they lack the coordination that would be seen with a true phase-conjugate mirror. Nevertheless, a fuzzy image of the pinhole forms on the screen as piece after piece of the wavefront arrives.

The retroreflectors in plastic sheets come in two types. One type consists of prisms embedded in the plastic. An arriving light ray is reflected from one of the internal faces of the prism and then from a second and possibly a third internal face. It travels out of the plastic parallel to the incident ray. The retroreflector is not perfect because the returned ray is displaced from the incident ray and because the section of the wavefront is inverted by the prism. (The retroreflector is so small that the section of wavefront can be regarded as straight.) The Reflexite Corporation manufactures this type of retroreflecting array. The prisms are clustered in hexagonal cells, each cell .15 millimeter across. Areas of the array that are between the cells do not take part in retroreflection.

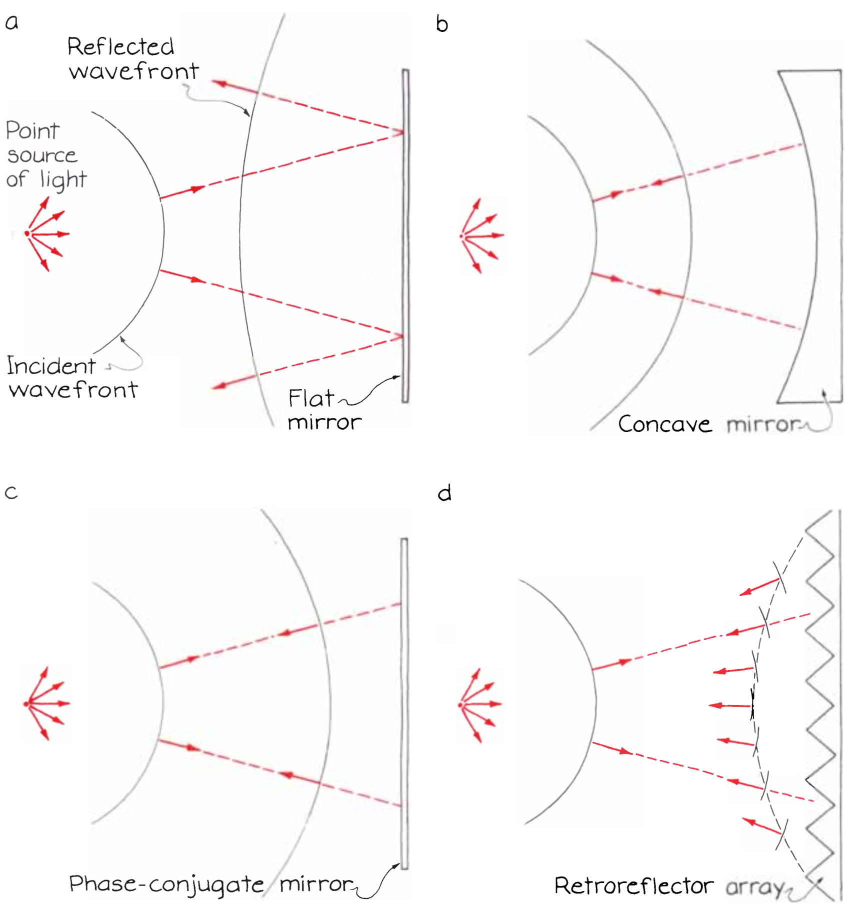
Scotchlite, made by the 3M Corporation, consists of glass beads embedded in a flexible backing. An arriving ray of light is refracted to the back of the bead, from which it is reflected back to the front. When it leaves the bead, it is parallel to the incident ray. This retroreflector is imperfect in that it displaces the final ray and inverts the final wavefront.

One novel application of retroreflectors involves removing the distortion of a beam of light that passes through a nonuniform material. Place a concave mirror in the path of the light from the pinhole, positioning the mirror to get the best image of the pinhole on the screen. Insert a distorting element in front of the mirror. The element could be one of those ruffled plastic sheets that normally cover fluorescent bulbs, a glass slide covered with airplane glue, the plastic "bubble" material used in packing or almost any other nonuniform, transparent material. The distorting element destroys the image of the pinhole, scattering the light into a complex pattern that fills the screen. If you shift the distorting element across the opening of the mirror, the pattern on the screen changes chaotically.

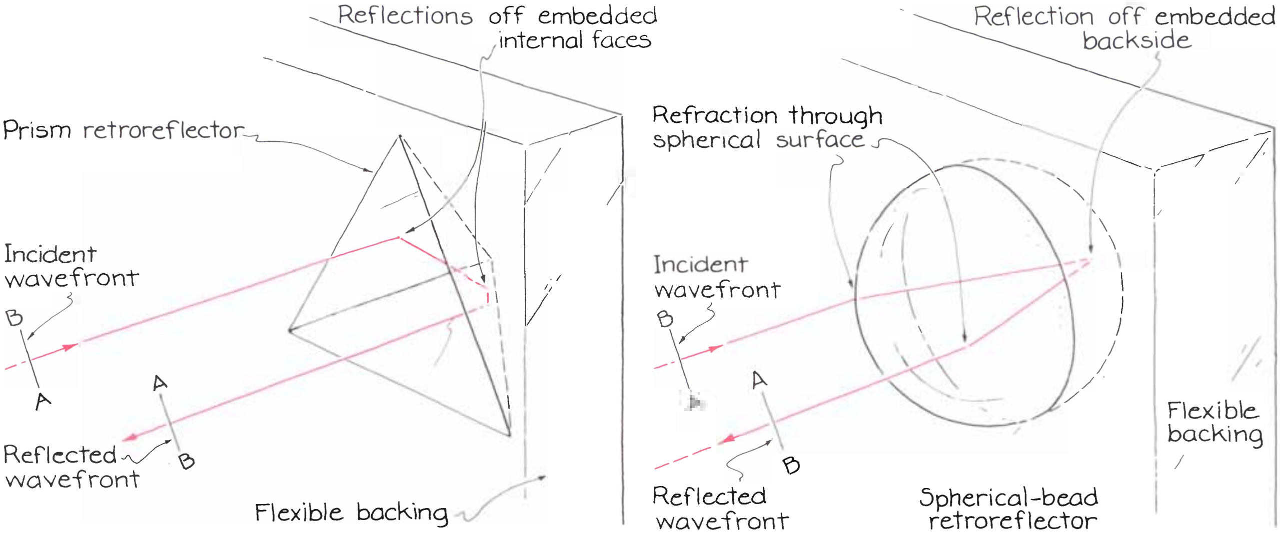
Now replace the mirror with a retroreflecting array. The exact position and orientation of the array are not important, but the distorting element should be directly in front of it and both the array and the distorting element should be illuminated. Although the distorting element jumbles the light and seemingly ruins any information about the pinhole's shape, an image of the pinhole appears on the screen. It is not perfect, being dim and fuzzy, but nonetheless it is there.

The image appears because the array almost exactly reverses the light rays emerging from the distorting element. The rays then travel back through almost the same distorting features they encountered on their first passage. The second passage reverses the distortion of each ray, the beam is reconstituted and the image of the pinhole is formed on the screen. Pepper suggests you can consider that the rays have been time-reversed by the array. If you move the distorting element across the array, the pinhole image changes little or not at all. The reason is that the time the light takes to travel from an irregular feature in the element to the array and back is so short that the feature is essentially stationary during the trip.

Note that the distortion is removed because the light passes through the distorting element twice. (If you place the element in front of the pinhole or the screen, the light passes through it only once.) The light rays must make their second pass through the element along approximately the same paths they followed on the first pass. Remember the retroreflectors return light rays that are slightly displaced from the incident rays. If the element and



*How different types of mirror reflect*



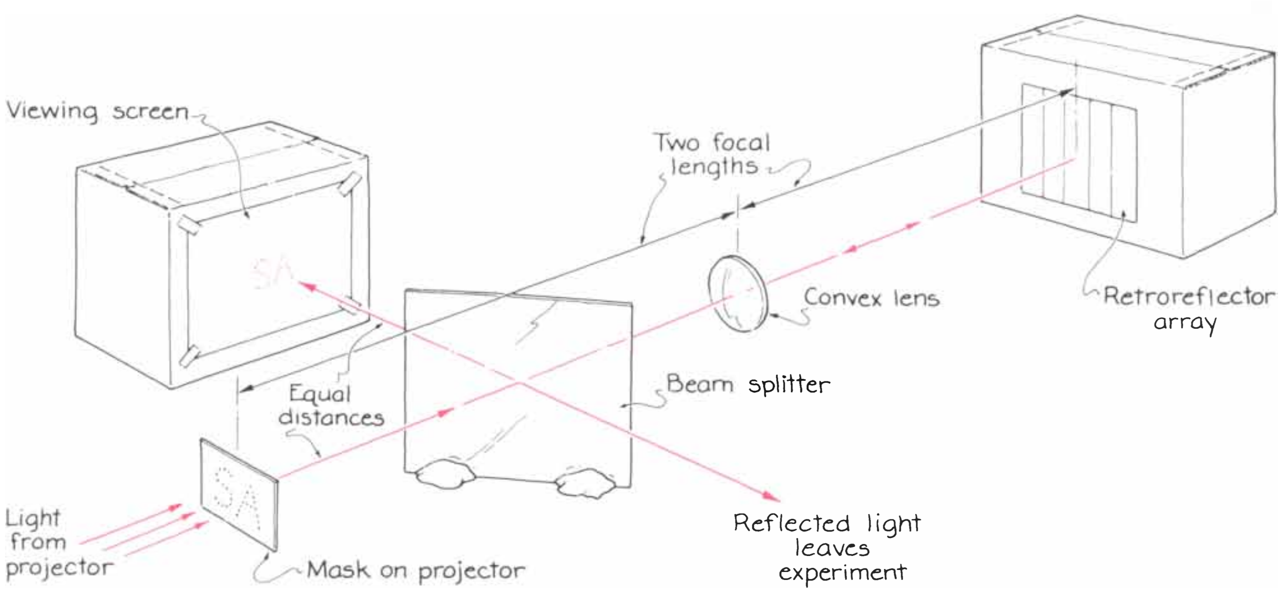

*Two types of retroreflector*

the array are separated too much, the returned rays encounter different distorting features on their second passage through the element and the distortion is not removed. A large separation may also mean that less light is intercepted and returned by the array. A dimmer image then results.

You can do similar demonstrations if you replace the projector and pinhole with a laser. The image is marred, however, by extra interference in the light. The interference gives rise to a complex pattern of bright and dark lines. Some of the interference is due to the diffraction of light by dust motes on the beam splitter or to nonuniformities in the glass. Some of the light reaching the screen has been distorted only once, and so it retains the distortion. The periodicity of the beads or hexagonal clusters on the retroreflector sheet also introduces interference in the light reaching the screen.

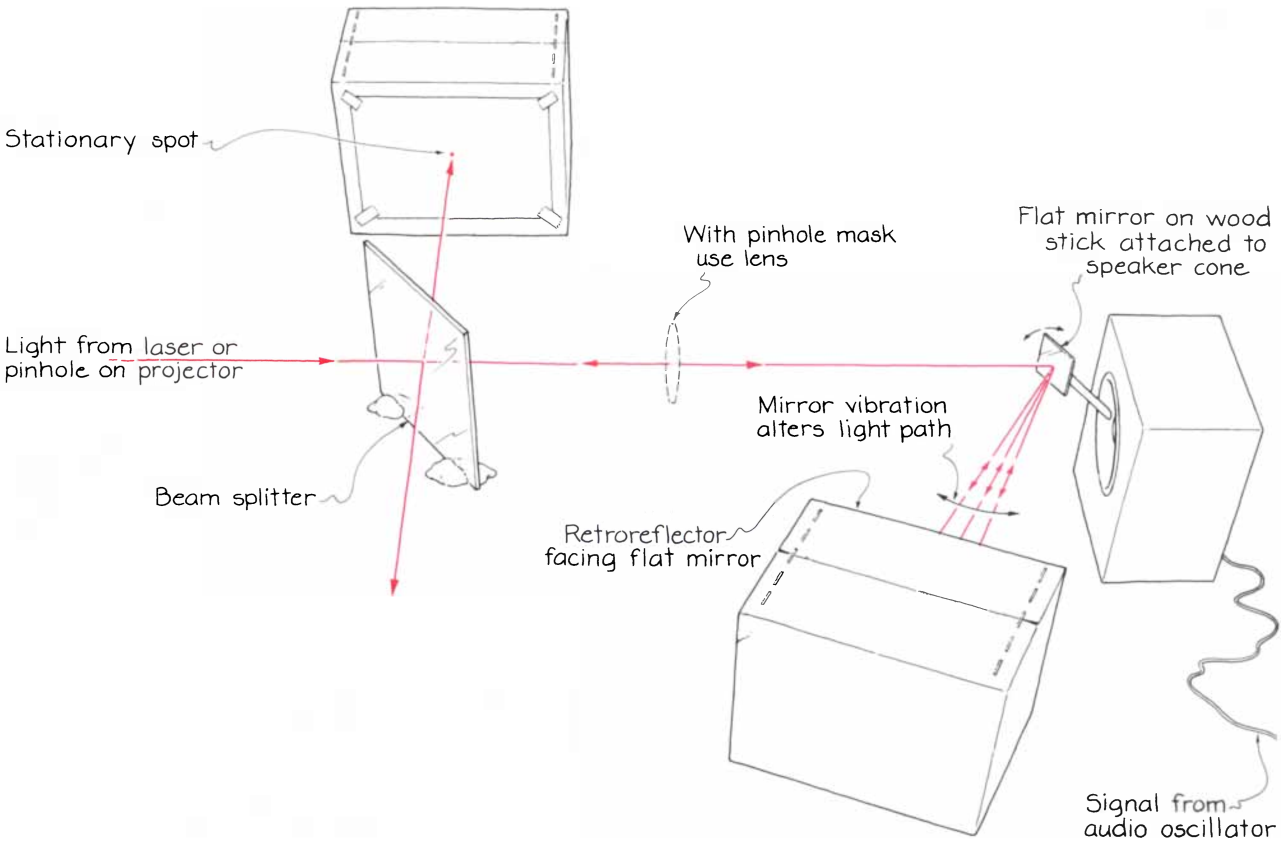
When you do this demonstration, you may be confused by an extra spot of light on the screen. Because of the imperfect construction of the retroreflecting array, part of the light returning from it is reflected in the same way as light is reflected from a flat mirror. If the plane of the array is perpendicular to the beam of light, this mirrorlike reflection reaches the screen and can be confused with the pinhole image formed by retroreflection. The extra spot of light is easy to identify because it shifts over the screen if you tilt the array with respect to the beam. You can remove the spot by erecting a cardboard folder between the array and the screen so that the laser beam just clears the edge of the folder. Then rotate the array until the mirrorlike reflection appears on the folder and is thus blocked from the screen.

Many experiments can be done with the laser, the retroreflecting array and common mirrors and lenses. I particularly enjoyed blowing smoke or throwing fine powder into the path of the laser beam in order to follow its travel. In one arrangement the beam passed through the beam splitter, was reflect-

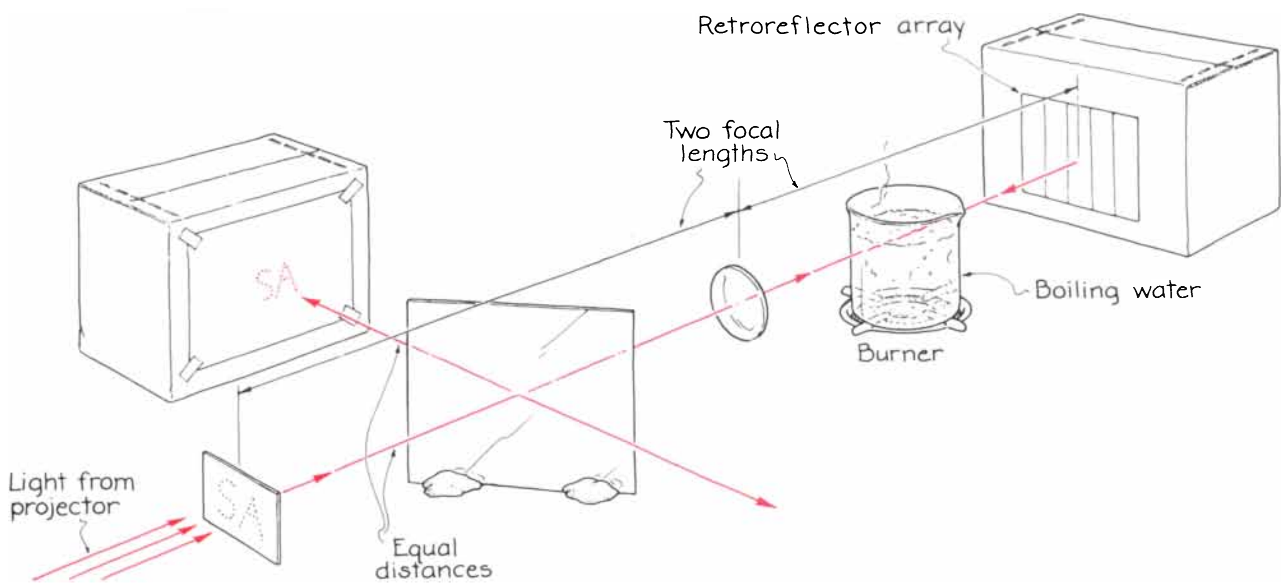



*Creating an image from a group of pinholes* ed from a tilted flat mirror and then was returned from a retroreflecting array. The trip demonstrates two distinct types of reflection: the familiar reflection from the mirror and the effect of retroreflection, which folds the beam back on itself.

I also enjoyed inserting diffraction gratings and other distorting elements in front of the array. In each case retroreflection removes the distortion imposed on the light in its first passage through the element. Some of the laser light, however, is reflected from the front surface of the distorting element and therefore never makes a second passage through it. This distortion is not removed. To block the mirrorlike reflection rotate the distorting element along with the array so that the reflection appears on the cardboard fold-




*Removing vibrational distortion from a beam*




*David M. Pepper's experiment with distortion in turbulence* er. What then remains on the screen is the light that was retroreflected from the array and made a second passage through the distorting element.

More than one pinhole can be imaged using the retroreflecting array. Mount on the front of the projector a mask bearing a series of pinholes arranged to form one or two small letters. All the experiments can be repeated with this design. For example, when the distorting element and the retroreflecting array are in position, a fuzzy image of the letters appears on the screen. Each pinhole creates its own image; the composite forms the letters on the screen.

To improve the image install a convex lens between the beam splitter and the retroreflecting array at a distance of two focal lengths from the pinholes. Place the array two focal lengths on the other side of the lens. Pepper says that the focal length of the lens is not critical but that the experiment works well when the length is between 10 and 30 centimeters. A simple magnifying glass will do.

To tune the alignment move the lens and the array until a sharp, inverted image appears on a sheet of paper held just in front of the array. Remove the paper and move the viewing screen until the image is as sharp as you can make it. A distorting object placed directly in front of the array makes the image dimmer and fuzzier, but it is still visible. A lens is needed in this demonstration to compensate for the slight spreading of the beam produced by the retroreflectors. With this arrangement you might try replacing the cardboard mask on the front of the projector. Remove the front lens and put an ordinary slide inside the projector.

Pepper has also investigated how a retroreflecting array can undo the distortion introduced by vibration. He glued a flat, lightweight mirror to a wood tongue depresser, which he attached to a speaker cone with double-sided adhesive tape. Then he illuminated the mirror with the light passing through the beam splitter. (The source of the light can be either a laser or a projector with a pinhole mask.) The light was reflected from the mirror onto a retroreflecting array. Next Pepper drove the speaker with the amplified signal from an audio oscillator. As the mirror vibrated, the angle at which the beam was reflected to the array changed continuously. Nevertheless, the screen still displayed a stationary image of the laser's exit aperture (or of the pinhole). Even when the mirror vibrated rapidly, it moved considerably slower than the speed of light. Hence it was effectively stationary while the light moved from the mirror to the array and back. When Pepper replaced the array with a flat mirror, the screen showed a smeared, kinetic pattern.

Pepper did a final demonstration that involved removing the distortion imposed by a turbulence. He set up the basic arrangement of equipment with a mask of pinholes on the projector and a retroreflecting array at one end of the optical path. He also placed a convex lens between the beam splitter and the mirror and installed an unlighted gas burner just below the optical path between the lens and the mirror. (It could have been between the lens and the beam splitter.) The pinhole, the viewing screen and the array were each two focal lengths from the lens. Pepper tuned the alignment of the system and lighted the burner.

One would expect the image to dance and flicker as the beam passed through the turbulent convection rising from the flame. It would do so with a flat mirror, but in this case it remained frozen in place. Although the distortion of the beam varied constantly in the turbulence, the array returned a clear image because the light gets through the disrupting features twice before they have time to change.

Pepper then placed a beaker of water on the burner and in the path of the light. As the water heated, the convection in it distorted the light beam, and again the retroreflecting array removed the distortion. When the water began to boil, the bubbles prevented some of the light from reaching the array. As a result the image on the screen became somewhat dimmer.

Here is a puzzle for you to solve. If the arrays are mirrors, why do you not see a retroreflection of your face when you look at an array in normal room illumination? (You will see your face if the room is brightly lighted and the array is oriented just right, but that reflection is due to imperfect construction of the array.)

You will probably find plenty of other demonstrations to do with the retroreflecting arrays. Pepper and I should like to hear about your experiments.

Last December I described a kaleidoscope with polarizing filters and a curved reflecting sheet inside the tube. This Karascope, which was commissioned by New York's Museum of Modern Art, was invented and patented by Judith Karelitz of New York City. It is for sale from the museum for $20.75 postpaid; the address of the museum is 11 West 53 Street, New York, N.Y. 10019. Please be sure to mark correspondence to the attention of the Mail Order Department.


# 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 __________

# 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


THE FOURTH DIMENSION SIMPLY EXPLAINED: A COLLECTION OF ESSAYS SELECTED FROM THOSE SUBMITTED IN THE SCIENTIFIC AMERICAN'S PRIZE COMPETITION. Edited by Henry P. Manning. Dover Publications, Inc., 1960.

FLATLAND: A ROMANCE OF MANY DIMENSIONS. Edwin A. Abbott. New American Library, 1984.

THE FOURTH DIMENSION: TOWARD A GEOMETRY OF HIGHER REALITY. Rudy Rucker. Houghton Mifflin Company, 1984.

THE SHAPE OF SPACE: HOW TO VISUALIZE SURFACES AND THREE-DIMENSIONAL MANIFOLDS. Jeffrey R. Weeks. Marcel Dekker, Inc., 1985.


## CAN THE U.S. TRUST THE U.S.S.R.?


COMPLIANCE WITH SALT I AGREEMENTS. U.S. Department of State Administration Report in *Department of State Bulletin,* Vol. 78, No. 2013, pages 10–14; April, 1978.

CHARGES OF TREATY VIOLATIONS: MUCH LESS THAN MEETS THE EYE. John Pike and Jonathan Rich in *FAS Public Interest Report,* Vol. 37, No. 3, pages 10–20; March, 1984.

A REVIEW OF REAGAN'S REPORT ON COMPLIANCE. Thomas K. Longstreth in *Arms Control Today,* No. 3, pages 6–9; March/April, 1984.

CIA IS SKEPTICAL THAT NEW SOVIET RADAR IS PART OF AN ABM DEFENSE SYSTEM. Michael Gordon in *National Journal,* Vol. 17, No. 10, pages 523–524; March 9, 1985.


## THE DYNAMICS OF PROTEINS


DYNAMICS OF FOLDED PROTEINS. J. Andrew McCammon, Bruce R. Gelin and Martin Karplus in *Nature,* Vol. 267, No. 5612, pages 585–589; June 16, 1977.

DYNAMICS OF PROTEINS. P. G. Debrunner and H. Frauenfelder in *Annual Review of Physical Chemistry,* Vol. 33, pages 283–299; 1982.

DYNAMICS OF PROTEINS: ELEMENTS AND FUNCTION. M. Karplus and J. A. McCammon in *Annual Review of Biochemistry,* Vol. 52, pages 263–300; 1983.

MAPPING PROTEIN DYNAMICS BY X-RAY DIFFRACTION. Dagmar Ringe and Gregory A. Petsko in *Progress in Biophysics and Molecular Biology,* Vol. 45, No. 3, pages 197–235; 1985.


## THE QUANTIZED HALL EFFECT


THEORY OF THE QUANTIZED HALL CONDUCTANCE. B. I. Halperin in *Helvetica Physica Acta,* Vol. 56, Nos. 1–3, pages 75–102; 1983.

THE QUANTIZED HALL EFFECT. H. L. Stormer and D. C. Tsui in *Science,* Vol. 220, No. 4603, pages 1241–1246; June 17, 1983.

PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON ELECTRICAL PROPERTIES OF TWO-DIMENSIONAL SYSTEMS, HELD AT OXFORD UNIVERSITY, 5–9 SEPTEMBER, 1983. Edited by R. J. Nicholas in *Surface Science,* Vol. 142; 1984.

VON KLITZING WINS NOBEL PRIZE FOR QUANTUM HALL EFFECT. B. Schwarzschild in *Physics Today,* Vol. 38, No. 12, pages 17–20; December, 1985.


## ISLAMIC ASTRONOMY


THE OBSERVATORY IN ISLAM AND ITS PLACE IN THE GENERAL HISTORY OF THE OBSERVATORY. Aydin Sayilli. Publications of the Turkish Historical Society, 1960.

THE SCIENTIFIC ENTERPRISE. A. I. Sabra in *The World of Islam: Faith, People, Culture,* edited by Bernard Lewis. Alfred A. Knopf, Inc., 1976.

STUDIES IN THE ISLAMIC EXACT SCIENCES. E. S. Kennedy. American University of Beirut, 1983.


## INTERNAL RHYTHMS IN BIRD MIGRATION


CIRCANNUAL SYSTEMS. E. Gwinner in *Handbook of Behavioral Neurobiology, Vol. 4,* edited by J. Aschoff. Plenum Press, 1981.

BIRD MIGRATION. Chris Mead. Facts on File Publications, 1983.

CIRCANNUAL RHYTHMS IN THE CONTROL OF AVIAN MIGRATIONS. E. Gwinner in *Advances in the Study of Behavior,* Vol. 16, pages 191–228; 1986.


## OPTICAL GYROSCOPES


THE LASER GYRO. Frederick Aronowitz in *Laser Applications, Vol. 1,* edited by Monte Ross. Academic Press, 1971.

MULTIOSCILLATOR LASER GYROS. W. W. Chow, J. B. Hambenne, T. J. Hutchings, V. E. Sanders, M. Sargent III and M. O. Scully in *IEEE Journal of Quantum Electronics,* Vol. QE-16, No. 9, pages 918–936; September, 1980.

THE RING LASER GYRO. W. W. Chow, J. Gea-Banacloche, L. M. Pedrotti, V. E. Sanders, W. Scleich and M. O. Scully in *Reviews of Modern Physics,* Vol. 57, No. 1, pages 61–104; January, 1985.

GYROSCOPES MAY CEASE SPINNING. Graham J. Martin in *IEEE Spectrum,* Vol. 23, No. 2, pages 48–53; February, 1986.


## THE "STRESS" OF BEING BORN


CATECHOLAMINE RELEASE IN THE NEWBORN. H. Lagercrantz and P. Bistoletti in *Pediatrics Research,* Vol. 11, No. 8, pages 889–893; August, 1977.

OF LABOUR AND THE LUNGS. R. E. Olver in *Archives of Disease in Childhood,* Vol. 56, No. 9, pages 659–662; September, 1981.

FETAL SCALP CATECHOLAMINES DURING LABOR. Peter Bistoletti, Lars Nylund, Hugo Lagercrantz, Paul Hjemdahl and Helge Ström in *American Journal of Obstetrics and Gynecology,* Vol. 147, No. 7, pages 785–788; December 1, 1983.

SYMPATHOADRENAL ACTIVITY AND PERIPHERAL BLOOD FLOW AFTER BIRTH: COMPARISON IN INFANTS DELIVERED VAGINALLY AND BY CESAREAN SECTION. G. Faxelius, H. Lagercrantz and A. Yao in *Journal of Pediatrics,* Vol. 105, No. 1, pages 144–148; July, 1984.


## PLAYING THE BAROQUE TRUMPET


THE MUSIC AND HISTORY OF THE BAROQUE TRUMPET BEFORE 1721. Don L. Smithers. Syracuse University Press, 1973.

ACOUSTICS AND THE PERFORMANCE OF MUSIC. Jürgen Meyer. Verlag Das Musikinstrument, Frankfurt, 1978.

REGENERATION IN BRASS WIND INSTRUMENTS. S. J. Elliott and J. M. Bowsher in *Journal of Sound and Vibration,* Vol. 83, No. 2, pages 181–217; July 22, 1982.


## THE AMATEUR SCIENTIST


RETROREFLECTIVE ARRAYS AS APPROXIMATE PHASE CONJUGATORS. H. H. Barrett and S. F. Jacobs in *Optics Letters,* Vol. 4, No. 6, pages 190–192; 1979.

EXPERIMENTS WITH RETRODIRECTIVE ARRAYS. Stephen F. Jacobs in *Optical Engineering,* Vol. 21, No. 2, pages 281–283; March/April, 1982.

# ˌän·trə·prə·ˈnər

**Entrepreneur**

*• A leader; e.g., with the innovative technology to produce ten straight years of earnings growth.*
*• One who organizes, manages and assumes the risks of a business or activity, in this case, one with 123,000 employees worldwide.*
*• One who serves as a middleman; e.g., between the demands of the market for excellence and the demands of the next decade for innovation.*

The demands of the market present a high-technology company with a maze of opportunities. Rockwell International responds by giving each of its operations what entrepreneurs have always needed. Freedom to move and act quickly. Rockwell operations apply advanced technology to create innovative quality products.

The result of this entrepreneurial philosophy: a leading role in five diverse areas of commercial and government business–contributing to overall corporate sales exceeding $11 billion.

To learn more about us, write: Rockwell International, Department 815S-4, 600 Grant Street, Pittsburgh, PA 15219.

**Rockwell International**

**...where science gets down to business**

**Aerospace / Electronics / Automotive**
**General Industries / A-B Industrial Automation**

As Grant Anderson is always quick
to point out, when you can throw the hammer
112 feet in Scotland's Highland Games, "You
don't need a winning smile to impress the judge."
The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
PRODUCT OF SCOTLAND
John Dewar & Sons Ltd
PERTH
SCOTLAND
86.8° PROOF
Authentic.
The Dewar Highlander
86.8 PROOF · © 1986 SCHENLEY IMPORTS CO., N.Y., N.Y.