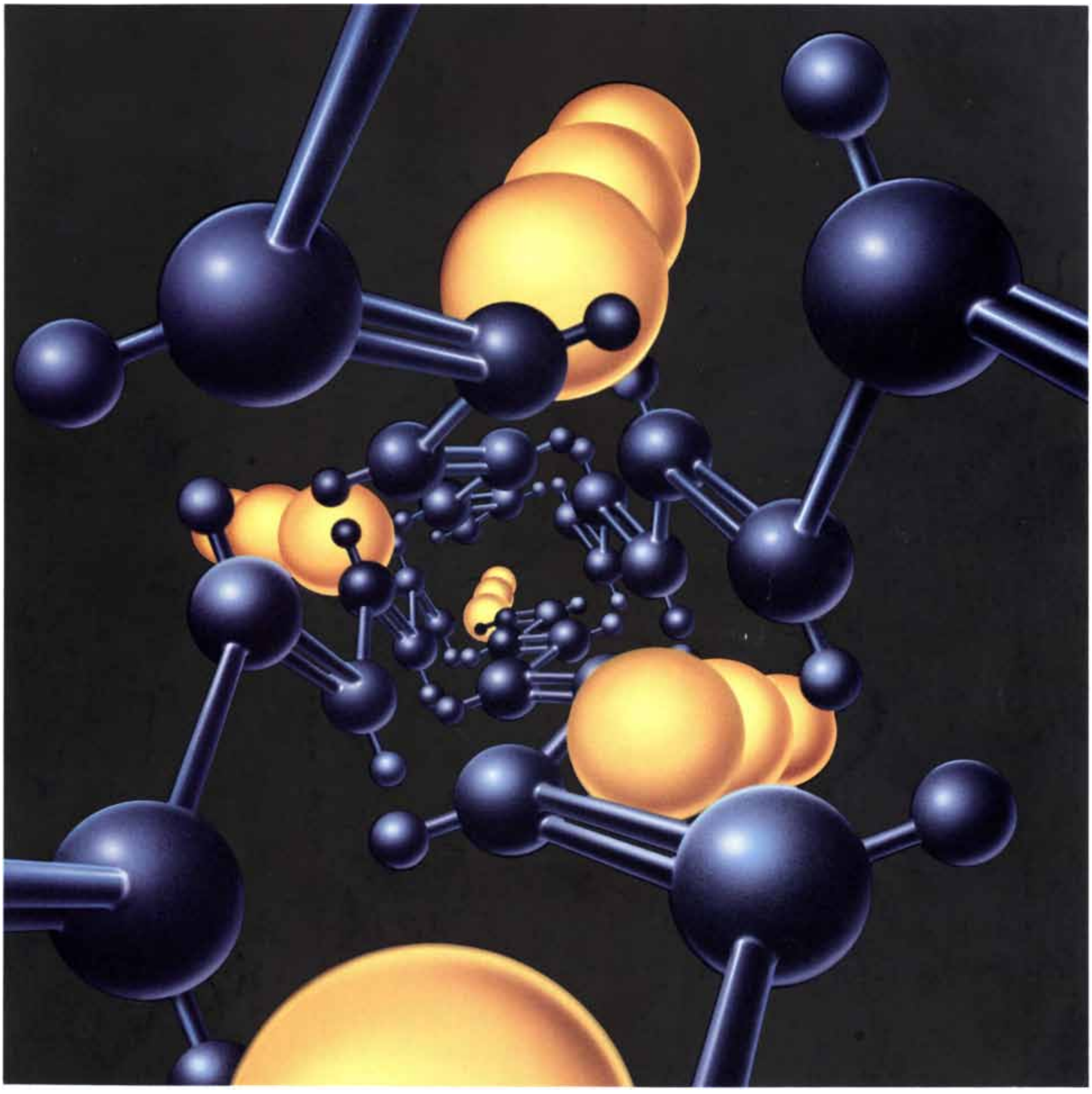
SCIENTIFIC AMERICAN

FEBRUARY 1988
$3.50

A slightly flawed cosmic mirror allows the universe to exist.
Why is Mars too cold, Venus too hot and the earth just right?
The opossum adapts to almost anything; it even eats rattlesnakes.



Plastic Conducts Electricity *as well as copper wire does when the carbon chains are doped with ionic triplets.*

It'll make you more c

There's no way out of traffic. There's often no way around it. But driving the Accord LX Sedan can certainly make the best of it.

The Accord is efficient and aerodynamic on the outside. It takes up no unnecessary space. Yet its ergonomically designed interior will accommodate five full-grown adults. And their personal space.

Its superior acoustical environment and controllable climate can turn the daily commute into a quiet, meditative time. Or a chance to put the electronic stereo/cassette system and fully powered amenities through their paces.

omfortable in crowds.

On the other hand, when you have the good fortune to hit the open road, the Accord's 2.0 liter, 12-valve engine redefines rush hour. Its 4-wheel double wishbone suspension holds on relentlessly. And that should serve to increase your comfort zone significantly.

Unless, of course, admiring glances make you feel uncomfortable.

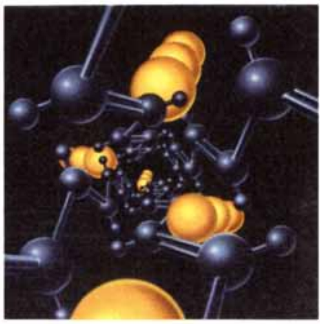
HONDA

The Accord LX

SCIENTIFIC AMERICAN

February 1988 Volume 258 Number 2

DEPARTMENTS



THE COVER image is a model of doped polyacetylene, a plastic that on the basis of volume conducts electricity better than copper (see "Plastics That Conduct Electricity," by Richard B. Kaner and Alan G. MacDiarmid, page 106). The yellow triiodide dopants are interspersed among four polyacetylene chains: carbon atoms, with hydrogens attached, linked by alternating single and double bonds. Undoped polyacetylene, like most plastics, is an insulator.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin

Page	Source
41	The Bettmann Archive
42–47	Joan Starwood
48	Edward Bell
51–56	Jerome Kuhl
83	Garland R. Marshall
84–85	Alan D. Iselin
86	David I. Gottlieb
87–89	Alan D. Iselin
90	National Aeronautics and Space Administration
91	National Aeronautics and Space Administration (*left*), The Hale Observatories (*right*)
92–94	George V. Kelvin
95	National Aeronautics and Space Administration
96–97	George V. Kelvin
99	James Kilkelly, courtesy of Savannah River Ecology Plant
100–104	Patricia J. Wynne
107	James Kilkelly
108-108D	Ian Worpole
109	Tokyo Institute of Technology
111	James Kilkelly
112A	John Moss, courtesy of The Royal Society of London
112B	Michael Goodman
113–114	Edward Bell
115	Edward Bell (*top*), Laurie Grace (*bottom*)
116	Edward Bell
117	John Moss, courtesy of The Royal Society of London
118–119	Scala/Art Resource
120	The Granger Collection
121–122	Bob Conrad
124	Al Satterwhite, The Image Bank
125–127	Michael Goodman
129	Andrew Christie
130	David Wiseman (*top*), Richard Dawkins (*bottom*)
131	Richard Dawkins

SCIENTIFIC AMERICAN

Established 1845

The Great Scot Parka, the cargo jean, and other "clothing that works" from Lands' End.

At Lands' End, we have a special affection for "clothing that works". Clothing that puts function first, fashion second. The kind we wear ourselves.

So we're forever adding our two cents' worth of suggestions on how various items can be improved in design and manufacture.

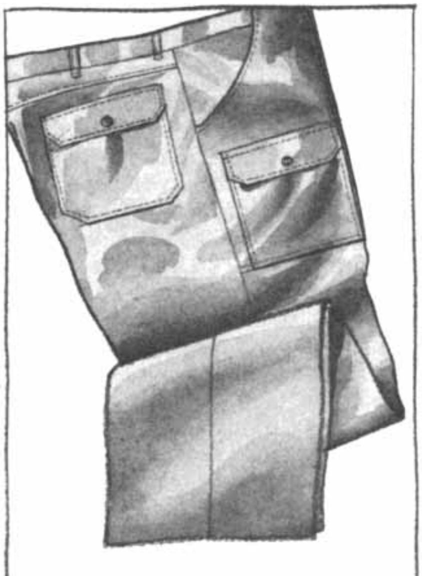
Every now and then, this practice really pays off, as it has in the development of our Great Scot Parka—arguably the most functional piece of outerwear you can buy.

Great Scot Parka $79.00

We give it all the features you expect in a mountain parka: weather-resistant shell, drawstring hood, drawstring waist, adjustable cuffs, storm flap. Features that allow you to seal out wet and cold. And then, in our usual Lands' End manner, we give it an extra feature you may not expect: a warm wool lining. A lining that extends even up into the hood, to keep your ears warm. All that done, we then give our Great Scot a price that represents significant value: just $79.00 ($69.00 for children).

The backpack of jeans.

Cargo room. That's the story with our 6-pocket Cargo Jean. You can fill the pockets with just about everything you need for a day's hiking or sailing: snacks, compass, blues harp, Proust. Even strap gear to the sturdy tunnel

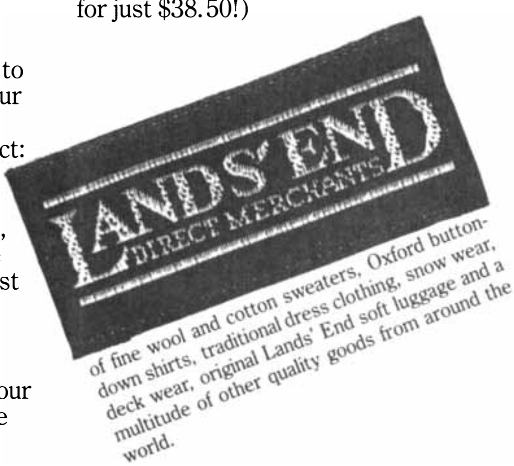


6-Pocket Cargo Jean $26.50

belt loops. This jean won't let you down. And it has the full, easy-moving cut that all Lands' End garments give you. The price? Just $26.50.

"Tough as the game" Rugby Shirts.

Not your flimsy sport shirts, but authentic, tough 100% cotton jerseys tested in rough and tumble action. Gussetted under the arms for easy movement, and equipped with rubber buttons that won't gouge you. Best feature? They're preshrunk, so they'll fit for as long as you wear them. Just $36.50. (And we now have XXL for just $38.50!)



"Tough as the game" Rugby Shirt $36.50

First quality, then price.

Along with "clothing that works", our catalog also offers all manner of traditional dress clothing, soft luggage, and other quality merchandise from around the world. Everything is direct-to-you, at prices that give new meaning to the word "value".

We offer you first, quality. Then, price. And always, always service.

Plus a guarantee so unconditional, we state it proudly in just two words:

GUARANTEED. PERIOD.®

If we're new to you, we don't ask that you trust us. Simply try us. Call toll-free 24 hours a day at 1-800-356-4444.

Or, fill in the coupon and let us send you a free copy of our Lands' End catalog.

Please send free catalog.
Lands' End Dept. Q-04
Dodgeville, WI 53595

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

Or call Toll-free:
1-800-356-4444

SCIENTIFIC AMERICAN

In Other Languages

LE SCIENZE
L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Via G. De Alessandri, 11
20144 Milano, Italy
Advertising correspondence:
Publietas, S.p.A., Via Cino de Duca, 5,
20122 Milano, Italy

サイエンス
Y880/copy Y9600/year Y13,000/(abroad)
Editorial, subscription, advertising correspondence:
Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y CIENCIA
450 Ptas/copy 4950Ptas/year $35/(abroad)
Editorial, subscription, advertising correspondence:
Prensa Cientifica S.A.,
Calabria, 235-239
08029 Barcelona, Spain

POUR LA SCIENCE
27FF/copy 265FF/year 315FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum DER WISSENSCHAFT
9.80 DM/copy 99 DM/year 112.20 DM/(abroad)
Editorial, subscription correspondence:
Spektrum der Wissenschaft GmbH & Co.
Moenchhofstrasse, 15
D-6900 Heidelberg,
Federal Republic of Germany
Advertising correspondence:
Gesellschaft Für Wirtschaftspublizistik
Kasernenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

科学
1.40RMB/copy 16RMB/year $24/(abroad)
Editorial, subscription correspondence:
ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

В МИРЕ НАУКИ
2R/copy 24R/year $70/(abroad)
Editorial correspondence:
MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.
Subscription correspondence:
Victor Kamkin, Inc.
12224 Parklawn Drive,
Rockville, MD 20852, USA

TUDOMÁNY
98Ft/copy 1,176Ft/year 2,100Ft/(abroad)
Editorial correspondence:
TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary
Subscription correspondence:
"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

العلوم
1KD/copy 10KD/year $40/(abroad)
Editorial, subscription, advertising correspondence:
MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 – Kuwait

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0550 Telex: 236115

LETTERS

To the Editors:

Charles H. Bennett's excellent article "Demons, Engines and the Second Law" [SCIENTIFIC AMERICAN, November, 1987] prompts me to suggest a feasible demon. It must be flawed, but I cannot see how.

Consider a hollow sphere with a fully reflective inner surface. A black ball is mounted at its center. All radiation from the ball is reflected back to the center, and so the ball's temperature stays constant.

Consider a hollow ellipsoid with a fully reflective inner surface. A black ball is mounted at each of its two foci. All radiation from each ball is reflected to the other ball, and so the temperatures of the two balls stay constant and equal.

Now consider a composite structure: a hollow ellipsoid cut away at one end, where it is joined to a hollow sphere, mounted so that its center is a focus of the ellipsoid. A black ball is mounted at each of the foci. All radiation from the focus within the ellipsoid is reflected to the other focus, but most of the radiation from that focus (the center of the sphere) is reflected back to the same point. Hence there is a net flow of energy from the first focus to the second.

TONY GREENFIELD

Whitwell, Hertfordshire
England

To the Editors:

The second law of thermodynamics, as Bennett correctly explains, is an expression of the extremely small probability of such events as the flow of heat from colder to warmer regions of a system. When the hypothetical event takes place on a macroscopic scale, the probability is so small that the second law becomes an assertion of impossibility. Events on smaller scales, however, generally have larger probabilities, until at the atomic level deviations from the average are the rule. At this level violations of the second law in its simplest form are not at all surprising, and mechanisms (call them "demons") that can take advantage of fluctuations to bring about "unnatural" changes would not conflict with any fundamental principle.

From this statistical point of view it is not necessary to disprove absolutely the existence of "demons"; it would seem more natural to try to calculate what they can accomplish. For example, the original demon, operating the sliding door between two chambers, may suffer from limitations of the kind described in the article, but I would have expected that long before these limitations came into play he would be in trouble. He would have progressively fewer chances of finding a hot molecule he could let pass into the desired chamber without an even hotter one's escaping. This difficulty would determine, at least roughly, how large a temperature difference the demon could produce in a given amount of gas, or, conversely, for how much gas a given temperature difference could be achieved.

LESTER GUTTMAN

Chicago, Ill.

To the Editors:

It seems to me that Bennett's ingenious modification of the Szilard engine [November, 1987, page 114] refutes his claimed exorcism of Maxwell's demon. Returning the key to the neutral position (erasing information about a molecule's location) involves exactly the same types of operations as are needed to record the information, and if the latter operation can be done with negligible work, then so can the former.

HARRY G. PARKE

Brooklyn, N.Y.

To the Editors:

In Bennett's very interesting article on the exorcism of Maxwell's demon the argument seems to be missing a step. It is stated that after Szilard's engine—a version of the demon—extracts work from the single molecule in the cylinder, all the work is necessary to reset the machine's memory. Although it is clear that resetting the memory does require work, it is not clear that it requires as much as is generated by the expansion in the cylinder. Indeed, it would seem that it does not. Let the cylinder contain two or three molecules. One waits until they are all on one side of the partition before closing the gate. The wait and the measurement cost nothing. The work derived from the expansion is greater than it was when only one molecule was in the cylin-

der, and yet resetting the memory requires no more work than before.

MICHAEL P. SULZER

Arecibo, P.R.

To the Editors:

A common criticism, against which I offer little defense, is that I employed circular reasoning, using the second law to prove the second law. Attempts to understand the second law have always been somewhat tautological. Their main virtue is that they build confidence in the consistency of the law and in the impossibility of various devices that at first sight seem able to violate it. Readers in fact produced a great many of these devices, usually without believing they really work.

Tony Greenfield's letter, for example, describes a specially shaped mirror-lined cavity that seems to direct radiation preferentially from one body (*A,* in the ellipsoidal cavity) to another (*B,* in the sphere), violating the second law. Even in a purely ellipsoidal cavity, however, a significant fraction of the radiation leaving one body would not go to the other but would miss it, undergo multiple reflections and return to the first focus. This occurs because the bodies at the foci are not points but have a finite radius. In the composite cavity, in which part of the ellipsoid is missing, even more radiation from *A* misses *B* and returns to *A*. In contrast, little or none (depending on the exact geometry of the sphero-ellipsoid) of the radiation from *B* undergoes multiple reflections. Almost all of it is reflected once off the ellipsoid to *A* or once off the sphere back to *B*. Analysis shows that the fraction of radiation from *A* that returns to *A* (by multiple reflections) exactly equals the fraction from *B* that returns to *B* (by a single reflection off the sphere), and the second law is satisfied.

Lester Guttman's letter notes that microscopic violations of the second law occur spontaneously and suggests therefore that no harm would be done if a demon could produce a small temperature difference in a small system containing only a few molecules. This is true only if the demon is not able to produce any difference in *average* temperature between two parts of the small system. If the demon were able to produce a difference in average temperature, the same difference could be conducted to two macroscopic bodies placed in thermal contact with the two parts of the small system, resulting in a serious macroscopic violation of the second law.

Harry Parke's letter suggests that returning the key to the neutral position should require no more work than the other operations shown on page 114, and similar suggestions have been made by other readers. In reality, returning the key does require more work because it is logically irreversible, compressing two states of the apparatus into one. All the other operations are logically reversible. Other readers suggest that the neutral state be omitted entirely and each new measurement be allowed simply to overwrite the value (L or R) of the old. In this scheme erasure per se is avoided, but the measurement step then becomes logically irreversible, not because of the acquisition of new information but because of concomitant destruction of old information.

Some readers proposed that by scaling up the entire Szilard engine or using a heavier molecule a greater work yield could be obtained, but this is not so. The work yield is independent of these parameters because a molecule's Brownian velocity varies in inverse relation to the square root of its mass, and the pressure of a gas varies inversely with the volume of the cylinder it fills.

Michael Sulzer's letter presents a more effective way to increase the work yield: put several molecules in the cylinder instead of only one, and wait for them all to collect on one side. Other readers rediscovered a related idea of Denis Gabor's. Gabor noted that by waiting for even one molecule to enter a small volume, trapping it there, and then allowing it to expand into a much larger volume, an arbitrarily large work yield (proportional to the logarithm of the volume ratio) could be obtained. Unlike the original Szilard engine, either of these "high performance" engines must wait for an improbable event before proceeding with its power stroke. It turns out that waiting itself has a thermodynamic cost that exactly offsets the extra work yield.

But why should it cost energy just to wait for a rare event? Could the waiting not be done by a passive device such as a mousetrap, which expends no energy until the desired event occurs, and then expends only some fixed amount, regardless of the duration of the wait? No: the rarer the event is, the more energy must be expended to trap it. This is so because a trap, like any other supposedly irreversible device, has some chance of running backward—of spontaneously opening up under the influence of Brownian motion and allowing the mouse (or molecule) it holds to escape. These spontaneous escapes become less frequent when the trap's spring is made stiffer, but they cannot be prevented altogether. For the high-performance Szilard engine to run forward as intended, the frequency of spontaneous escape must be made less than the frequency of successful trapping, and this in turn requires a spring so stiff that the entire increased work yield of the engine must be spent to recock it.

Several readers proposed devices that would convert the molecule's motion into a weak, random reciprocating motion of a piston rod. This random motion in one dimension, however, is no more useful mechanically than the three-dimensional random Brownian motion of the molecule, from which it is derived.

CHARLES H. BENNETT

To the Editors:

The short item on the toxic properties of potato skins ["Science and the Citizen," SCIENTIFIC AMERICAN, September, 1987] prompts a reply from the Irish point of view. The potato has of course been a staple of Irish rural diet since Stuart times. The poor husbandman must often have come to regret failing to earth up his spuds and thereby allowing tubers near the surface to be exposed to light.

The poisonous nature of green potatoes is well known on this island, and woe betide the cook who fails to excise the green areas when preparing potatoes for the pot. In general, however, potatoes are often boiled in their jacket and served that way, if they are of suitable quality.

The following children's skipping rhyme (to be recited in brogue, of course) testifies to the Irish predilection for potato skins:

"Where d'ye come from?"
"Donegal."
"How are the spuds there?"
"Big an' small."
"How d'ye eat them?"
"Skins an' all."
"Do they not harm you?"
"Not at all!"

JOHN PHILLIPS

Belfast
Northern Ireland


© 1987 BASF Corporation
Create a new world where the heart can help treat itself.
In North America, heart disease is claiming more lives than any other cause.
At BASF, we looked at the problem in new ways and saw the heart's muscle cells as an ally. The result was a BASF breakthrough: a cardiovascular "wonder drug." No single solution to heart disease exists. But by enlisting the body's own mechanisms, this BASF breakthrough is helping many cardiac patients lead healthier, more productive lives.
In one industry after another, from health care to cosmetics, our broad-based technologies help us create new worlds by seeing in new ways.
The Spirit of Innovation
BASF

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

FEBRUARY, 1938: "In the limestone stream bed of Purgatory River in southeastern Colorado, the positive record of a Jurassic day has been uncovered. There the almost perfect trails of both carnivorous and herbivorous dinosaurs have been preserved side by side. In addition to the large round tracks of a quadruped, the limestone contains numerous impressions from the feet of ancient three-toed bipeds. The entire assortment of ancient impressions, apparently, may have been made in the same distant day—possibly within the space of a single hour."

"When Mother Nature slips a cog and sends cold weather where warm weather ought to be, fruit growers must supply artificial heat to protect their valuable groves. Old-type smudge pots, long used in Florida and California, are being replaced in some sections by oil-burning heaters from which warm air is blown over large areas by a huge engine-driven propeller."

"After 26 years of constant research, the vitamin preventive of the disease beri-beri has been isolated, its chemical constitution determined and the vitamin itself synthesized at a cost far lower than that of recovering it from bran."

"The mystery of the endocrines is not yet wholly bared, but there is already much that can be definitely stated concerning these tiny organs that pour their secretions directly into the blood stream. We know that our physical growth, mental energy and general attitude toward life depend largely upon the concerted activity of a few glands which together do not weigh more than three-quarters of a pound!"

"Before New Year's Day of the year 2000, man may quite possibly reverse his present militant campaign against insects and actually strive to save many of the species he is now indiscriminately destroying with a grim determination. The dilemma arises through the very efficiency of modern methods of insect destruction, particularly such wholesale barrage effects as the use of airplanes for laying down of poison dusts by the square mile. These wide swathes of death sweep down not only the few insect species that are man's enemies but also the many that are his friends, or at most are merely neutral and harmless."

SCIENTIFIC AMERICAN

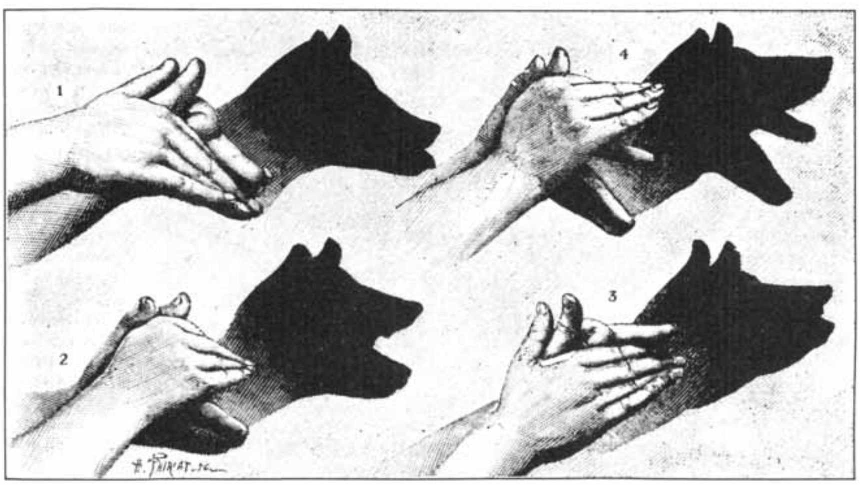
FEBRUARY, 1888: "The indefatigable de Lesseps has been forced to take a backward step in the construction of the transisthmian canal. Finding the work far from complete and his capital approaching exhaustion, he has determined to construct the canal for the present with locks. He has engaged Eiffel of Paris, the builder of the great 250-meter tower of the coming exposition, to construct for him a series of iron locks. These he proposes to place upon the line of the canal, and for the present to carry vessels across the high ground by means of the locks. This is by no means to be interpreted as an abandonment of his original project of a canal without locks."

"If the condensed breath collected on the cool window panes of a room where a number of persons have been assembled is burned, a smell as of singed hair will show the presence of organic matter; and if the condensed breath is allowed to remain on the windows for a few days, it will be found, on examination by a microscope, to be alive with animalcules."

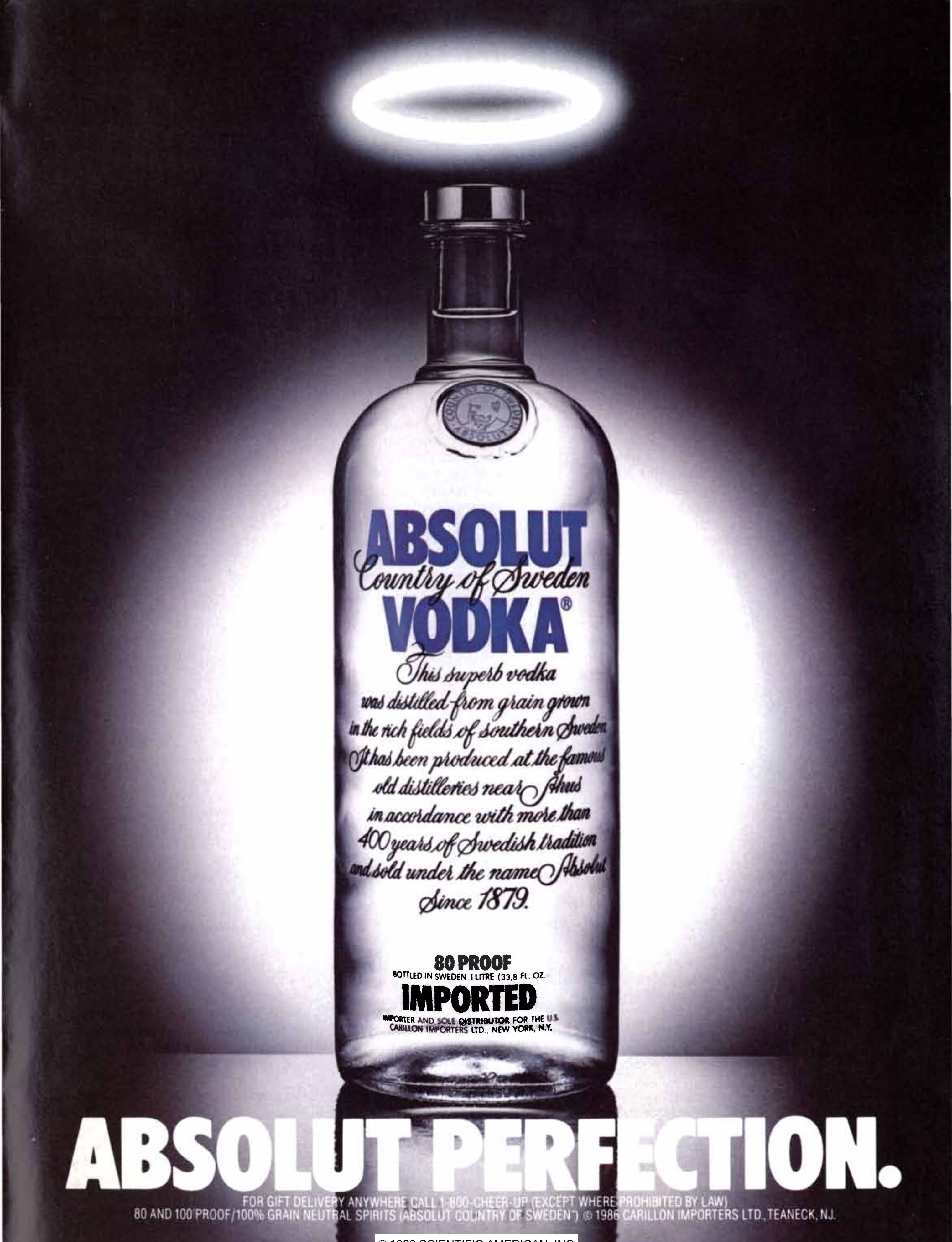
"For several years the problem of steam locomotion on ordinary roads has entered a new sphere. Instead of the heavy and cumbersome traction or road engines which are without speed and are used principally for moving merchandise, small 'automotive' vehicles have been substituted, which, being lighter and of greater speed, may be used for the transportation of travelers and for pleasure journeys."

"The distances of the stars are ascertained in the same manner as those of the sun and planets; that is, by parallax. Instead, however, of taking two stations at different parts of the earth's surface and laying down a base line between them, we take the diameter of the earth's orbit, or 183,000,000 miles, as the base, the observations being made at intervals of six months. Even with this immense line, however, the parallax is so small that it can only be detected by the most careful observations and accurate instruments. The star Alpha Centauri is closest to the earth."

"The idea of projecting the shadows of different objects (among others, the hands) upon a plane surface is very ancient, and it would be temerity to attempt to assign a date to the creation of these animals, which have served to amuse children in the evening since time immemorial. Within the past few years these rude figures have been improved, and the play of shadows has now become a true art instead of being a simple diversion."



A dog scents a piece of meat (1), *snaps at it* (2), *swallows* (3) *and yawns luxuriously* (4)


ABSOLUT
Country of Sweden
VODKA®
This superb vodka was distilled from grain grown in the rich fields of southern Sweden. It has been produced at the famous old distilleries near Åhus in accordance with more than 400 years of Swedish tradition and sold under the name Absolut Since 1879.
80 PROOF
BOTTLED IN SWEDEN 1 LITRE (33,8 FL. OZ.)
IMPORTED
IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.
CARILLON IMPORTERS LTD., NEW YORK, N.Y.
ABSOLUT PERFECTION.
FOR GIFT DELIVERY ANYWHERE CALL 1-800-CHEER-UP (EXCEPT WHERE PROHIBITED BY LAW)
80 AND 100 PROOF/100% GRAIN NEUTRAL SPIRITS (ABSOLUT COUNTRY OF SWEDEN™) © 1986 CARILLON IMPORTERS LTD., TEANECK, N.J.

THE AUTHORS

RAY WHITE and JEAN-MARC LALOUEL ("Chromosome Mapping with DNA Markers") are investigators in the Howard Hughes Medical Institute at the University of Utah. White is also cochairman of and professor in the department of human genetics at the University of Utah School of Medicine and director of the university's cystic fibrosis research center. He earned a Ph.D. in molecular biology from the Massachusetts Institute of Technology in 1971, did postdoctoral work in biochemistry at Stanford University and was on the faculty of the University of Massachusetts Medical School in Worcester before moving to Salt Lake City. White has been a leader in the effort to develop a complete map of the human genome. Lalouel, who grew up in Africa, is professor of human genetics at the University of Utah School of Medicine. He has an M.D. and a doctorate in quantitative genetics from the University of Paris, where he also completed training in psychiatry. He spent four years as professor and head of the department of human biology at his alma mater before joining the Howard Hughes Institute.

ROBERT K. ADAIR ("A Flaw in a Universal Mirror") is associate director for high-energy and nuclear physics at the Brookhaven National Laboratory. He was awarded his doctorate in physics in 1951 by the University of Wisconsin, where he remained as an instructor until 1953. He then spent six years at Brookhaven, leaving there in 1959 for Yale University. In almost 30 years at Yale he was professor, chairman of the physics department and director of the division of physical sciences. From 1978 to 1983 he was the editor of *Physical Review Letters.* Adair, who returned to Brookhaven in 1987, is physicist to baseball's National League.

DAVID I. GOTTLIEB ("GABAergic Neurons") is professor of neurobiology at the Washington University School of Medicine in St. Louis. He received a bachelor's degree in 1964 at Harpur College of the State University of New York (now S.U.N.Y. at Binghamton) and a doctorate in neurobiology from Washington University in 1972. To complement his training in neurobiology, he then stayed in St. Louis as a postdoctoral fellow in the department of biological chemistry. Gottlieb began work at the medical school in 1976.

JAMES F. KASTING, OWEN B. TOON and JAMES B. POLLACK ("How Climate Evolved on the Terrestrial Planets") are colleagues at the Ames Research Center of the National Aeronautics and Space Administration. Kasting, who is a research scientist there, earned a bachelor's degree in chemistry and physics at Harvard College in 1975 and a doctorate in atmospheric science from the University of Michigan in 1979. He was a postdoctoral fellow at the National Center for Atmospheric Research and then at Ames until 1983, when he accepted his current position. Toon has been an associate fellow at Ames since he got a Ph.D. from Cornell University in 1975. He is one of the originators of the nuclear-winter theory and is the author of a hypothesis that clouds of nitric acid form in the Antarctic stratosphere, contributing to the development of the ozone "hole" there. Pollack is a research scientist at Ames. He holds a master's in physics from the University of California at Berkeley (1962) and a doctorate in astronomy from Harvard University (1965). He was a research physicist at the Smithsonian Astrophysical Observatory and then a senior research associate at Cornell University's Center for Radiophysics and Space Research before going to NASA.

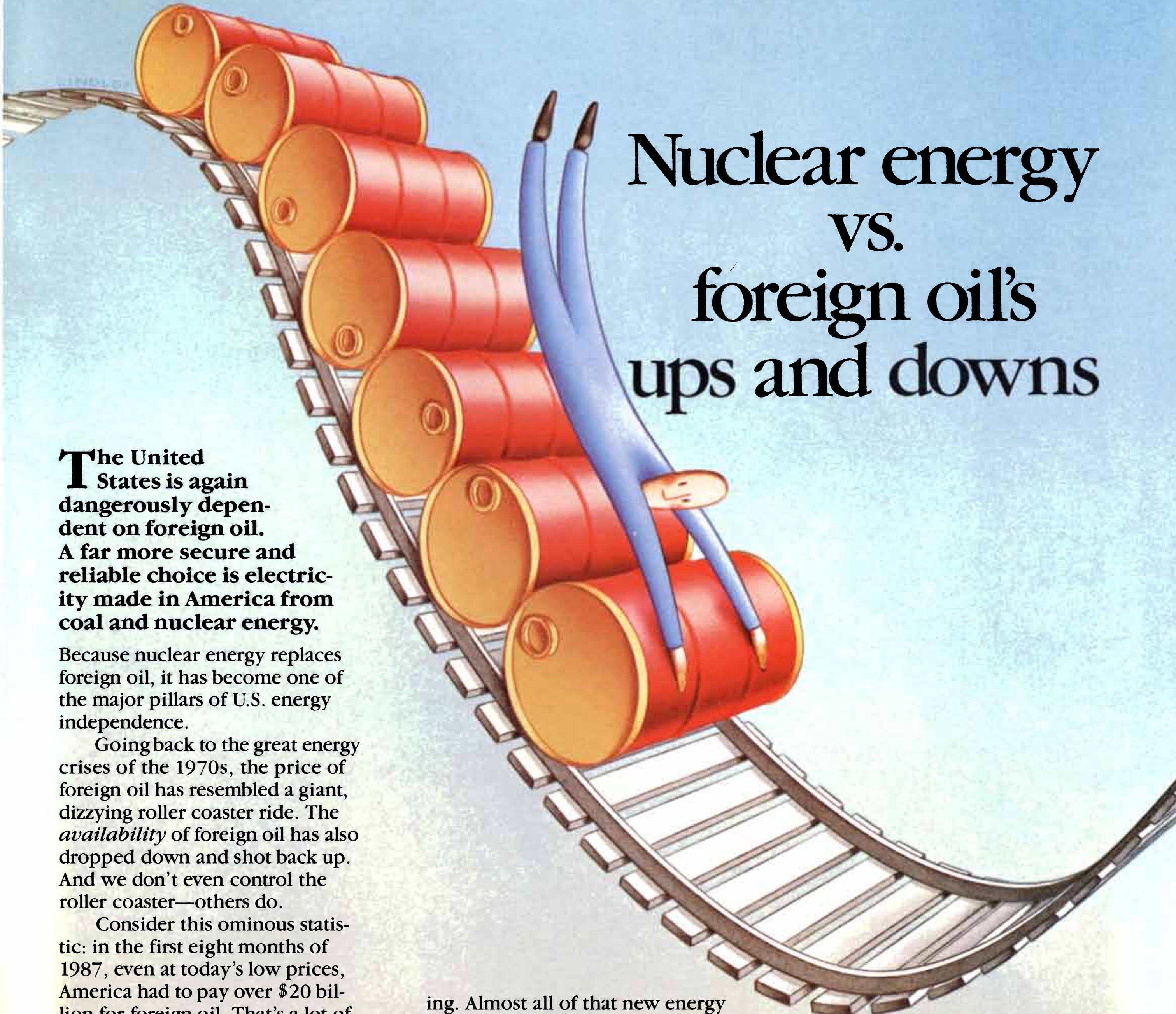
STEVEN N. AUSTAD ("The Adaptable Opossum") is assistant professor of organismic and evolutionary biology at Harvard University. Austad had a 1969 bachelor's degree in English from the University of California at Los Angeles when he became interested in animal behavior while working as a trainer in Hollywood. He thereupon got a bachelor's degree in biology at the California State University at Northridge (1976) and a Ph.D. in behavioral ecology from Purdue University (1981). He was a visiting instructor at Purdue and at the University of New Mexico before going to Harvard in 1986. Austad is married to a veterinarian and says he has a "houseful of mended and mending animals," including a dachshund with a pacemaker.

RICHARD B. KANER and ALAN G. MACDIARMID ("Plastics That Conduct Electricity") worked together when Kaner was a student in MacDiarmid's research group. Kaner is assistant professor of chemistry at the University of California at Los Angeles. He received his doctorate in inorganic chemistry from the University of Pennsylvania in 1984 and then did further study at the University of California at Berkeley. He took his current job in 1987. Kaner's research interests center on synthetic solid-state chemistry. MacDiarmid is professor of chemistry at Pennsylvania. Born in New Zealand, he obtained a bachelor's degree there in 1948 and a doctorate in chemistry from the University of Cambridge in 1955. He then taught at St. Andrews University, leaving there for Pennsylvania in 1964. Pioneering studies by MacDiarmid and his colleagues resulted in the synthesis, doping and characterization of polyacetylene, the first example of a conducting polymer.

JONATHAN M. BORWEIN and PETER B. BORWEIN ("Ramanujan and Pi") are brothers and natives of St. Andrews, Scotland. Jonathan Borwein, the elder, is professor of mathematics at Dalhousie University in Nova Scotia. He got his bachelor's degree in 1971 at the University of Western Ontario, where his father is still head of the mathematics department. He received his doctorate in mathematics in 1974 from the University of Oxford as an Ontario Rhodes Scholar and immediately joined the faculty of Dalhousie. The elder Borwein has also taught at Carnegie-Mellon University and has been a visiting researcher at the universities of Cambridge, Limoges and Montreal. Peter Borwein is associate professor of mathematics at Dalhousie. He has a bachelor's degree from the University of Western Ontario (1974) and a Ph.D. from the University of British Columbia (1979). The younger Borwein spent 1979–80 as a NATO research fellow at Oxford and has been at Dalhousie since then.

COLIN MCEVEDY ("The Bubonic Plague") is psychiatric consultant at St. Bernard's Hospital on the outskirts of London. The son of a surgeon, he qualified as a doctor himself in 1955, after studying at the University of Oxford. He worked at the Institute of Aviation Medicine while in the Royal Air Force and then trained in psychiatry at Maudsley Hospital and at Middlesex Hospital. McEvedy, who has a 1971 doctorate in medicine from Oxford, is a history buff and has published several historical atlases.

Nuclear energy vs. foreign oil's ups and downs



The United States is again dangerously dependent on foreign oil. A far more secure and reliable choice is electricity made in America from coal and nuclear energy.

Because nuclear energy replaces foreign oil, it has become one of the major pillars of U.S. energy independence.

Going back to the great energy crises of the 1970s, the price of foreign oil has resembled a giant, dizzying roller coaster ride. The *availability* of foreign oil has also dropped down and shot back up. And we don't even control the roller coaster—others do.

Consider this ominous statistic: in the first eight months of 1987, even at today's low prices, America had to pay over $20 billion for foreign oil. That's a lot of dollars leaving this country, adding to an already huge trade deficit.

Nuclear energy cuts oil imports

Clearly, the more energy we use in the form of electricity from coal and nuclear energy, the less oil we have to import.

Nuclear-generated electricity has already saved America over three billion barrels of oil, with billions more to be saved before the turn of the century. That's why it's so important for our energy self-reliance.

More electricity for a growing economy

Our economy needs plenty of *new* electrical energy to keep on growing. Almost all of that new energy is coming from coal and nuclear electric plants.

The truth is that nuclear energy is an everyday fact of life in the U.S. It's been generating electricity here for nearly 30 years. Throughout the country are more than 100 nuclear plants, and they are our second largest source of electric power. As our economy grows, we'll need more of those plants to avoid even more dependence on foreign oil.

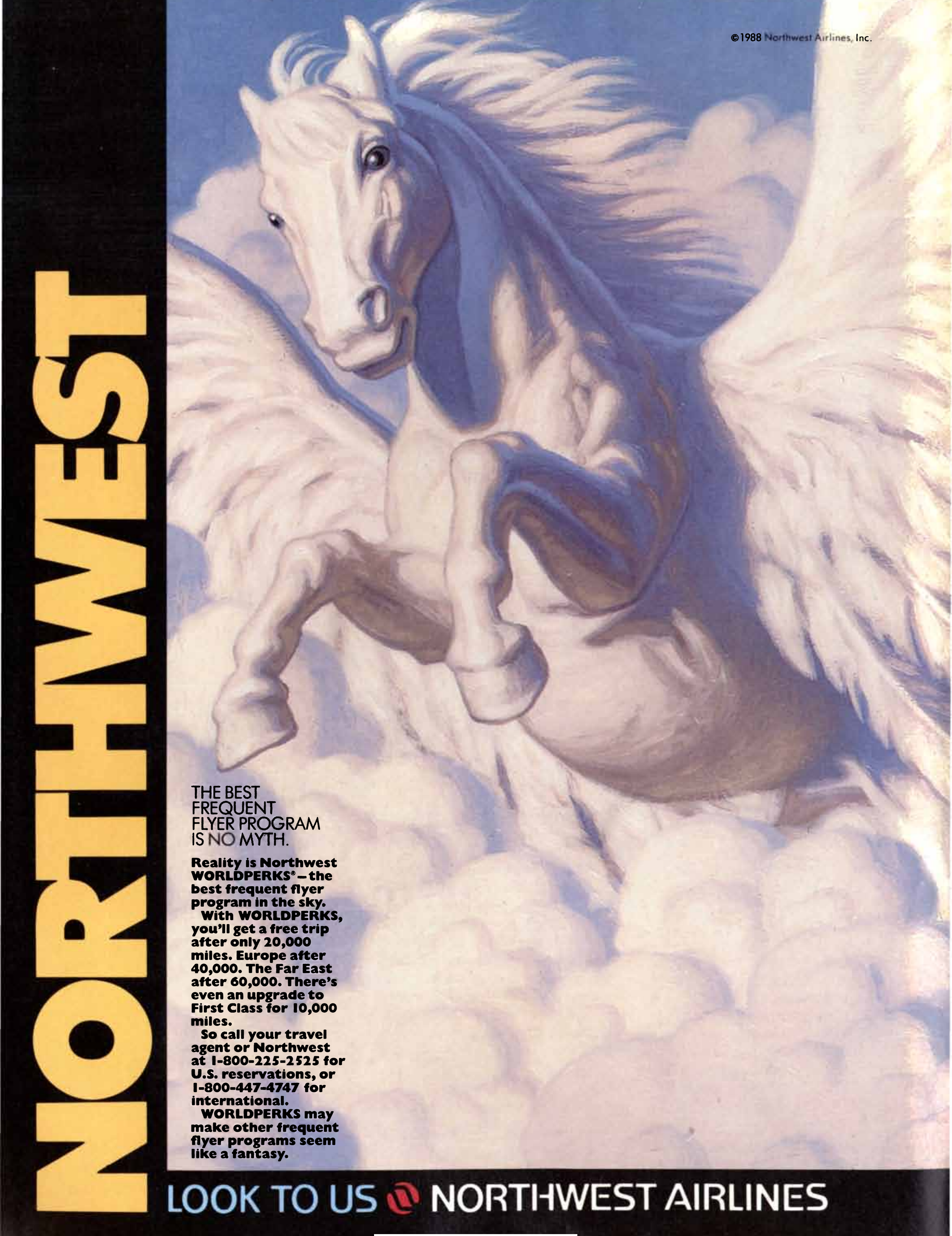
Safe energy for a secure future

Most important, nuclear energy is a safe, clean way to generate electricity. U.S. nuclear plants have a whole series of multiple backup safety systems to prevent accidents. Plus superthick containment buildings designed to protect the public even if something goes wrong. (It's a "Safety in Depth" system.)

The simple fact is this: America's energy independence depends in part on America's nuclear energy.

For a free booklet on energy independence, write to the U.S. Council for Energy Awareness, P.O. Box 66103, Dept. RR10, Washington, D.C. 20035. Please allow 2-3 weeks for delivery.

Information about energy America can count on
U.S. COUNCIL FOR ENERGY AWARENESS


©1988 Northwest Airlines, Inc.
NORTHWEST
THE BEST FREQUENT FLYER PROGRAM IS NO MYTH.
Reality is Northwest WORLDPERKS®—the best frequent flyer program in the sky.
With WORLDPERKS, you'll get a free trip after only 20,000 miles. Europe after 40,000. The Far East after 60,000. There's even an upgrade to First Class for 10,000 miles.
So call your travel agent or Northwest at 1-800-225-2525 for U.S. reservations, or 1-800-447-4747 for international.
WORLDPERKS may make other frequent flyer programs seem like a fantasy.
LOOK TO US NORTHWEST AIRLINES

SCIENCE AND THE CITIZEN

Science at the Summit

An informal U.S.-Soviet college has a role in arms control

The Washington summit highlighted the increasing importance of a group of civilian Soviet scientists who have become key arms-control advisers to General Secretary Mikhail S. Gorbachev. They include Roal'd Z. Sagdeyev, head of the Soviet Space Research Institute in Moscow, and Yevgeniy P. Velikhov, vice-president of the Academy of Sciences of the U.S.S.R., both of whom helped to found the Committee of Soviet Scientists against the Nuclear Threat after President Reagan's "Star Wars" speech of March, 1983. Both attended the summit, and some observers, including Marshall D. Shulman of Columbia University, maintain that scientists have never before had such influence with the Soviet leadership.

The Soviet scientists have also had a long-standing relationship with U.S. colleagues. Among the U.S. organizations that have maintained contacts even when government relations were frosty is the National Academy of Sciences (NAS). The academy's Committee on International Security and Arms Control, chaired by Wolfgang K. H. Panofsky, former head of the Stanford Linear Accelerator Center, has met its Soviet counterpart regularly for five years. The meetings are private and no negotiations take place, but both sides are debriefed by government officials.

"I think Gorbachev is very dependent on this new group of scientists," says Frank von Hippel of Princeton University and the Federation of American Scientists, an organization that also maintains informal Soviet liaisons. The existence of such contacts, von Hippel points out, enabled Velikhov to arrange for the Natural Resources Defense Council to monitor chemical explosions at the Soviet nuclear test site at Semipalatinsk; he also supervised a visit by U.S. congressmen to a disputed radar installation at Krasnoyarsk in Siberia. Kenneth L. Adelman, until recently director of the U.S. Arms Control and Disarmament Agency, takes a dimmer view: he charges that Velikhov is "polemical" in negotiations and is motivated by a desire to acquire U.S. "Star Wars" technology.

Yet the sounds as well as the recent action of Soviet science have been dovish. At the Washington summit Sagdeyev publicly discussed eliminating Soviet superiority in conventional arms in Europe and suggested possible 95 percent cuts in strategic nuclear weapons. Gorbachev called for a new joint science committee; such ideas appear to reflect Soviet eagerness to escape the pressures of the arms race in order to address economic issues.

That theme also ran through a meeting, held during the summit at the NAS, between Soviet academics and a group of U.S. scientists and industrialists. Participants heard Abel G. Aganbegyan, a chief architect of *perestroika,* discuss plans to make science—which is now "seriously behind the demands of life"—the main source of Soviet economic growth.

The warmness of *glasnost* has brought several civilian scientific collaborations into bud over the past two years. Joint workshops between the two science academies, halted in protest at Andrei Sakharov's internal exile, resumed recently with a meeting on condensed-matter theory held in December at the University of California at Santa Barbara; workshops on lasers and photochemistry, dense plasmas, earthquake prediction and vaccine production are planned. The NAS has proposed other collaborations, including participating in a Soviet program to put a million computers in Soviet schools. Frank Press, president of the NAS, notes that the cooperative ventures involve Soviet achievements "at or near the leading edge of science" and so yield benefits for both participants.

One such area is global climate and environmental change, where a joint government study is being planned. In space science, discussions are planned on joint tracking arrangements for a Soviet mission to Mars's moon Phobos, a possible Soviet Mars lander and the U.S. Mars Observer, as well as for a landing mission to Venus. Sagdeyev repeated in Washington his call for a joint manned mission to Mars, including a pointedly diplomatic offer to launch sealed, U.S.-guarded payloads.

At times, however, the chill of cold war intrudes. In a high-level medical meeting held in Washington last April, resuming a collaboration interrupted by the invasion of Afghanistan, U.S. representatives protested the appearance in Soviet publications of charges that the AIDS virus was created at Fort Detrick, Md., and insisted that such reports be stopped as a condition of collaborative study of the disease. Officials report that the charges have not been repeated since last August.

What is good for science is good for business. Several science-based U.S. companies have begun to capitalize on *glasnost* and *perestroika.* According to Howard A. Schneiderman, director of research at the Monsanto Company, his firm has increased milk yields by more than 25 percent in Soviet trials of genetically engineered bovine somatotropin. He says Monsanto is looking into a major joint effort to enhance the yield and other characteristics of crop plants by means of genetic engineering. —*Tim Beardsley*

Paleolithic Compassion

Did tender loving care help a Stone Age dwarf to survive?

As the Ice Age waned in Europe 11,000 years ago, the lot of humankind remained harsh. Groups roamed the countryside in search of berries, nuts, fish, game—or anything else that was edible. They hunted and prepared their food with tools made of stone, and they slept in caves. Survival of the fittest was the order of the day; the infirm and the aged, unable to fend for themselves, ordinarily perished quickly. This at


LANDAU MODEL

least is the picture painted by the bulk of the fossil record.

Two skeletons unearthed from the foot of the Italian peninsula add an element of tenderness to this stark tableau. The remains were found buried together inside the so-called Romito cave, near the modern city of Cosenza, in 1963 but were analyzed in detail for publication only recently by a team of four investigators. They report in *Nature* that one skeleton, called Romito 1, was a rather ordinary, although somewhat slight, 35-year-old female. Romito 2, however, was quite extraordinary: male, about 17 years old, and just over three feet tall—a dwarf. Carbon dating indicates that the bones are 11,150 years old. The finding, according to David W. Frayer of the University of Kansas, the lead investigator, extends by more than 5,000 years the earliest incidence of dwarfism known. More important, Frayer says, the finding reveals that Paleolithic hunters and gatherers could "extend sympathy to someone who never contributed much to the society."

11,000-YEAR-OLD DWARF exhumed in southern Italy had short, bowed forearms, a symptom of acromesomelic dysplasia.

Frayer and his colleagues found Romito 2 had been "severely deformed" by a rare kind of dwarfism—which inbreeding makes more likely—known as acromesomelic dysplasia. His forearms were extremely short and bowed and he probably moved them at the elbows only with difficulty. Walking with an awkward "waddling" gait, he would have tired quickly, particularly in the mountainous terrain of southern Italy.

Frayer says the dwarf's burial in the cave, which seems to have had some ceremonial importance, indicates that "he wasn't discriminated against; he was accepted." Engravings of animals decorate the cavern walls, and it overlooks a valley that was probably an important passageway for Upper Paleolithic groups.

The middle-aged female, Romito 1, may have played a special role in keeping the little male alive, according to Frayer. The workers who originally excavated the skeletons reported that the dwarf's skull rested on the female's cheek and that her arm embraced his body. Frayer notes that similarities between the two skeletons—for example, the cranial bones of both are joined in similar patterns—indicate the female may have been the dwarf's mother.

Other anthropologists think Frayer may overestimate the degree to which the dwarf depended on others. "Certainly some support mechanism had to be there," Donald J. Ortner of the Smithsonian Institution says. "But it's quite remarkable how capable people are at adapting to physical disabilities of one sort or another." Erik Trinkaus of the University of New Mexico points out that the fossil record suggests Paleolithic hominids could survive with broken arms, arthritis and other handicaps. But no hunter-gatherer society, Trinkaus asserts, has been known to nurse back to health a member with a severely broken leg or other immobilizing debility. "The individual would just be left behind," he says. *—John Horgan*

As California Goes...

An unselective carcinogen law spreads alarm in industry

If it is true that as California goes, so goes the nation, U.S. industry had better watch out. Beginning on March 1, a "minimum list" of chemicals that are—or might be—human carcinogens will be subject to stringent new state controls under Proposition 65, a voter initiative passed in November, 1986. Proposition 65 will make it illegal to contaminate any source of drinking water with a chemical on the minimum list; it also requires that products for sale containing one of the listed chemicals be accompanied by a warning. A product can be exempted only if its manufacturer can offer proof that it poses "no significant risk."

Proposition 65 states that the minimum list of carcinogens should include probable or possible carcinogens identified by two authoritative institutions, the National Toxicology Program (NTP) and the International Agency for Research on Cancer (IARC). A governor's scientific advisory committee can add other substances as necessary; each becomes controlled a year after its inclusion on the list.

Proposition 65's sweeping powers—including a bounty provision for whistle-blowers—were born of environmentalists' frustration with the pace of Federal efforts to control carcinogens. A recent report by the congressional Office of Technology Assessment concluded, for example, that no regulatory action has been taken on 23 out of 61 chemicals for which the NTP found strong evidence of carcinogenicity.

Indeed, the IARC and the NTP between them have identified about 235 chemicals as possible or probable human carcinogens; for the great majority the evidence is from animal studies. Usually tests in more than one species produced cancer at some dosage. These studies do not, however, evaluate how dangerous a carcinogen might be to human beings.

Industry complains that depending on lists of possible or probable human carcinogens for regulation is bad policy, since such lists will include substances that differ enormously in their potential for harm. James R. Fouts of the NTP concurs, saying its list of potential carcinogens was designed to be "an information vehicle to stimulate inquiries as to whether these chemicals should be regulated, not...*the* list of chemicals to be regulated." John Higginson of the Georgetown University Medical Center, a former director of the IARC, notes that the IARC's studies were "never meant to be more than a signal that something might be a human carcinogen."

Not Rolls Royce. Not Mercedes. No other luxury sedan.

Only the all new Chrysler New Yorker protects the entire car for 5 years or 50,000 miles. Just take care of normal upkeep. We cover the rest. No deductibles. No ifs, ands or buts.

With Chrysler's new Crystal Key Owner Care Program, for 5 years or 50,000 miles,* all you do is take care of normal maintenance, adjustments and wear items. Chrysler takes care of everything else. Engine, powertrain, air conditioning, steering, rust, suspension, electrical. Everything right down to the door handles.

And you're still protected on the engine and powertrain for 7 years or 70,000 miles and outer body rust-through for 7 years or 100,000 miles.*

And if you have a question or need assistance, there's a 24-hour toll-free hotline.

The Crystal Key Program starts with the all-new Chrysler New Yorker. Luxury. Elegance. Front-wheel drive. Fuel-injected V-6. Everything. Completely protected.

How does the competition stack up? They don't.

	BASIC CAR WARRANTY covers entire car except normal upkeep	MAJOR COMPONENTS PROTECTION	DEDUCTIBLE YOU PAY FOR REPAIR of major components	ENGINE PROTECTION†	POWER-TRAIN PROTECTION†	OUTER BODY RUST-THROUGH PROTECTION	24-HOUR TOLL-FREE HOTLINE
CHRYSLER NEW YORKER LANDAU	5 years/ 50,000 miles	5 years/ 50,000 miles	None	7 years/ 70,000 miles	7 years/ 70,000 miles	7 years/ 100,000 miles	Yes
ROLLS ROYCE CORNICHE	3 yr/ unlimited	3 yr/ unlimited	None	3 yr/ unlimited	3 yr/ unlimited	3 yr/ unlimited	No
MERCEDES BENZ	4 yr/ 50,000 miles	4 yr/ 50,000 miles	None	4 yr/ 50,000 miles	4 yr/ 50,000 miles	4 yr/ 50,000 miles	Yes
CADILLAC BROUGHAM	1 yr/ 12,000 miles	5 yr/ 50,000 miles	$100 after 1 yr/12,000 miles	6 yr/ 60,000 miles	6 yr/ 60,000 miles	6 yr/ 100,000 miles	No
LINCOLN TOWN CAR	1 yr/ 12,000 miles	6 yr/ 60,000 miles	$100 after 1 yr/12,000 miles	6 yr/ 60,000 miles	6 yr/ 60,000 miles	6 yr/ 100,000 miles	Yes
OLDS REGENCY BROUGHAM	1 yr/ 12,000 miles	3 yr/ 36,000 miles	$100 after 1 yr/12,000 miles	6 yr/ 60,000 miles	6 yr/ 60,000 miles	6 yr/ 100,000 miles	No
BUICK ELECTRA PARK AVE.	1 yr/ 12,000 miles	3 yr/ 36,000 miles	$100 after 1 yr/12,000 miles	6 yr/ 60,000 miles	6 yr/ 60,000 miles	6 yr/ 100,000 miles	No

See copies of these limited warranties at their dealers. Restrictions apply.

INTRODUCING CHRYSLER'S CRYSTAL KEY PROGRAM

Chrysler 7/70

Division of Chrysler Motors

CHRYSLER. DRIVING TO BE THE BEST.

*See limited warranty at dealer. Restrictions apply. Excludes Turbo model. †Deductibles may apply.

The governor of California, George Deukmejian, initially proposed to control only proved human carcinogens—of which there are fewer than 30—under Proposition 65. His advisory committee, however, decided during the past year to accept evidence from animal studies and has added most of the IARC- and NTP-identified carcinogens to the list.

Notwithstanding, in December the committee antagonized the proponents of Proposition 65 by recommending that any company in compliance with Federal statutes should be assumed to have met the "no significant risk" test automatically. Al Meyerhoff of the Natural Resources Defense Council says that the recommendation would "eviscerate" the statute, and that it would be challenged in court if it were accepted by the governor.

Proposition 65 seems likely to become the first broad-ranging statute controlling carcinogens (as well as substances that produce birth defects or cause sterility) that does not take into account exposure levels or degree of potency. Such indiscriminate regulation could backfire, Higginson fears, vitiating the warning effect through repetition. Some health officials believe overall risk might be limited more effectively by first strictly controlling exposure to substances that appear to be strong carcinogens, particularly in the workplace, where exposures are often dangerously high.

Defenders of Proposition 65's approach insist that the public must be warned of any evidence of carcinogenicity and that industry must prove safety. Carl Pope of the Sierra Club intends to campaign to get equivalents of Proposition 65 on the ballot in other states. *—T.M.B.*

PHYSICAL SCIENCES

A Light Matter

The usual is inverted: matter is diffracted and reflected by light

The unreality of the quantum world has become decidedly more real. According to quantum mechanics, light behaves as if it were composed of particles in some situations and of waves in others. The basic constituents of matter also exhibit particlelike or wavelike properties, depending on how observations are interpreted. This fundamental similarity in the nature of both matter and light is made strikingly clear by three recent experiments in which beams of atoms have been diffracted or reflected by light in much the same way as light is ordinarily diffracted or reflected by material crystals, beam splitters and mirrors.

The manipulation of atomic beams by light rests on the so-called gradient force that arises from the interaction of neutral atoms with a gradient, or change over distance, in the intensity of an oscillating electric field. The electric field component of a beam of laser light can be made to exert such a force, particularly if the beam is reflected back in the direction from which it came. The counterpropagating beams thereby interfere constructively to form a standing light wave: a periodic array of intensity peaks and valleys.

Taking advantage of the stable and strong electric field gradients of such a standing light wave, a group at the Massachusetts Institute of Technology reports in *Physical Review A* that it managed to scatter a beam of sodium atoms crossing the standing light wave at an angle of 90 degrees. By carefully controlling the angle of intersection as well as the shape and the wavelength of the standing light wave, the group showed that the beam could be scattered in only a few discrete directions.

Using a similar setup, a group from the École Normale Supérieure and the Collège de France in Paris has also observed a beam of atoms interacting with light in much the same way as it would with matter. The group explains in *Physical Review Letters* that it caused atoms in a beam to travel along the intensity valleys of a standing light wave just as an atomic beam "channels" itself between rows of atoms in a crystal.

Finally, a team based at the Institute of Spectroscopy in the U.S.S.R. reveals in *Journal of Experimental and Theoretical Physics Letters* that it was able to reflect an atomic beam off a "mirror" of light. The mirror consisted of an electric field gradient that extended just beyond a plate of quartz in which laser light was internally reflected. The workers found that the atomic beam could be reflected if it grazed the evanescent field at the surface of the plate.

Each group has plans to build on its experimental results. The Soviet investigators intend to construct a concave light mirror that could focus an atomic beam as tightly as possible, to a spot whose dimensions are comparable to the beam's wavelength. The M.I.T. group points out that a standing light wave preserves the phase of an atomic beam's "matter waves" as it scatters the beam into two or more beams. By recombining the scattered beams it is possible to create interference patterns that could be applied to measure extremely small shifts in their respective phases.

Perhaps the most intriguing idea is the French team's suggestion that atoms might conceivably be trapped at the intersection of three perpendicular standing light waves. Such a three-dimensional array of standing light waves can be envisioned as a "crystal" of light, since it is analogous to the three-dimensional array of atoms that constitutes a crystal of matter. *—Gregory Greenwell*

Cosmic Mirage

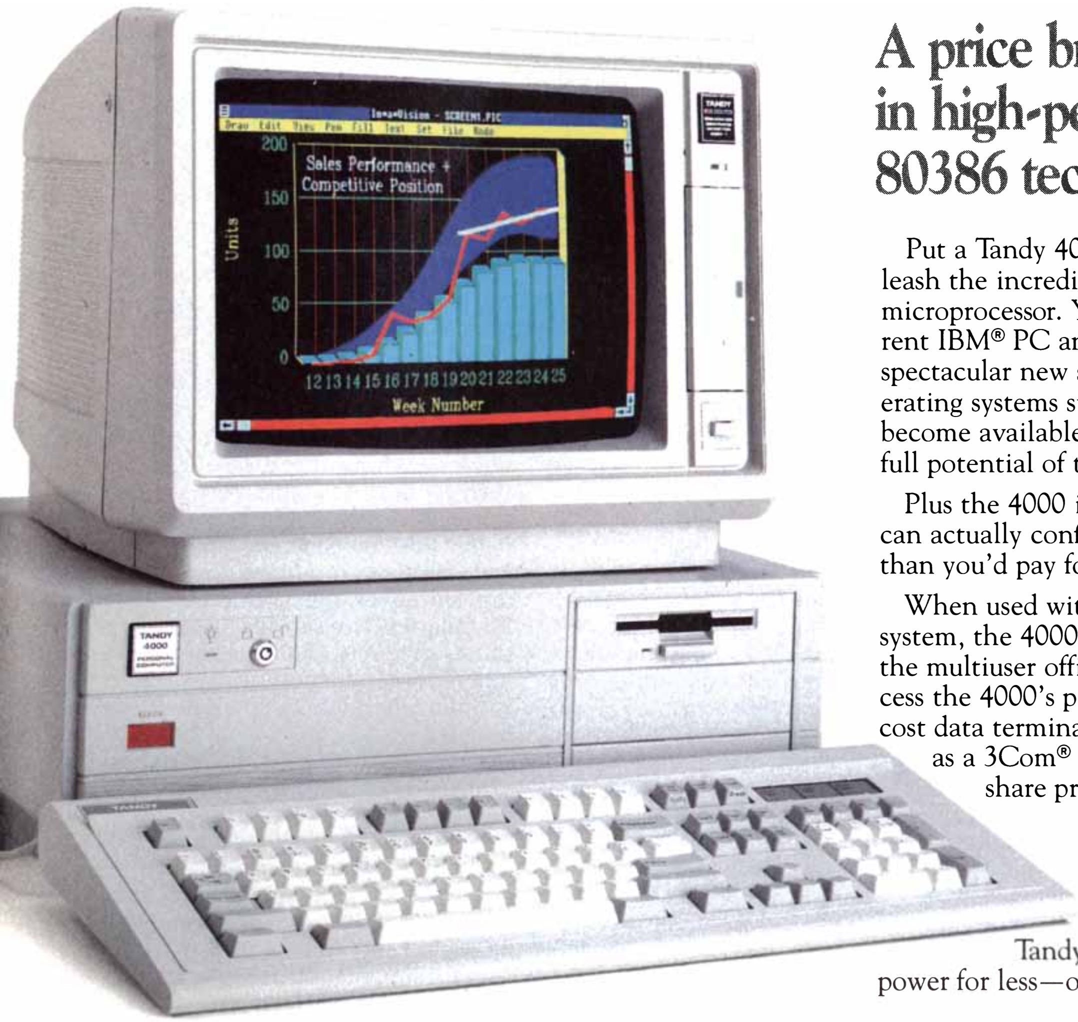
Gravitational lenses may have created vast "arcs" in space

Early last year two astronomers announced that they had found two huge starry arcs so geometrically perfect that they seem to have been drawn by a giant compass. The arcs appear in two separate clusters of galaxies, known as Abell 370 and cluster 2244–02, each some seven billion light-years from the earth. Both arcs measured more than 300,000 light-years long, or more than three times the span of the Milky Way, making them the longest continuously luminous objects ever seen. Vahe Petrosian of Stanford University and C. Roger Lynds of the National Optical Astronomy Observatories, who announced the discovery, speculated that the arcs might represent the shock front of a supernova or of stars being sucked from a galaxy by a black hole. They conceded, however, that no theory seemed to adequately explain the objects' unusual, precise shape.

Petrosian and Lynds now think the arcs are optical illusions: the images of single galaxies magnified to prodigious proportions by powerful gravitational lenses. They reached this conclusion after analyzing spectral data from the Abell 370 arc. The red shift of the arc's spectrum, consid-

Tandy Computers: Because there is no better value.™

The Tandy® 4000



A price breakthrough in high-performance 80386 technology.

Put a Tandy 4000 on your desk and unleash the incredible power of the 80386 microprocessor. You'll be able to run current IBM® PC and AT® software with spectacular new speed. And when new operating systems such as Microsoft® OS/2 become available, you'll be able to tap the full potential of the 80386.

Plus the 4000 is so cost effective, you can actually configure a 386 system for less than you'd pay for a competitor's 286 model.

When used with the XENIX® operating system, the 4000 can become the heart of the multiuser office system. Users can access the 4000's processing power from low-cost data terminals. Or configure the 4000 as a 3Com® workgroup file server to share programs and data files.

For data-intensive uses, accounting and financial planning, and even desktop publishing, the Tandy 4000 delivers incredible power for less—only $2599. (25-5000).

Send me a 1988 computer catalog.

Mail To: Radio Shack
Dept. 88-A-911
300 One Tandy Center
Fort Worth, TX 76102

Name ____________________

Company ____________________

Address ____________________

City ____________________

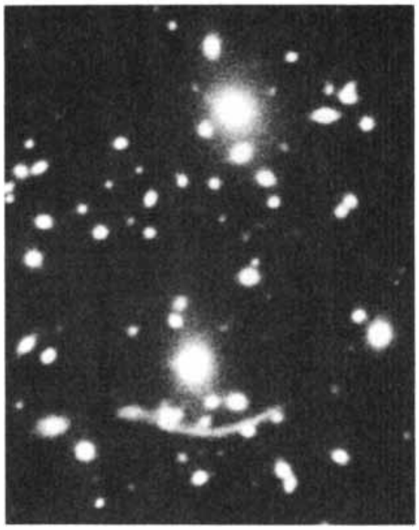
State ____________________

ZIP ____________________

Phone ____________________

Radio Shack®
COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

Price applies at Radio Shack Computer Centers and participating stores and dealers. Monitor, display adapter and operating systems sold separately. IBM and AT/Reg. TM and XT/TM IBM. XENIX/Reg. TM Microsoft Corp. 3Com/Reg. TM 3Com Corp.



VAST ARC is thought to be the image of a distant galaxy magnified by two closer galaxies (the brightest objects in this photograph) within the Abell 370 cluster.

ered a rough measure of its distance from the earth, is almost twice as large as the red shift of galaxies in Abell 370. The investigators conclud-ed that the light in the arc emanates not from the vicinity of Abell 370 but from a region almost twice as far away. Corroborating data have been reported by astronomers at the Toulouse Observatory in France.

Lynds believes a spectral analysis of the light in the 2244-02 arc will yield similar results. "I would be greatly surprised," he remarks, "if the two arcs don't reflect the same phenomenon."

Frederic H. Chaffee of the Smithsonian Institution's Mount Hopkins Observatory, an expert on gravitational lenses, finds the new explanation of the arcs' origin "quite convincing," and he adds, "When I heard of it, I felt a little foolish that I didn't think of it." He notes that the arcs represent the best examples ever observed of a so-called Einstein ring: the annular image that should appear around a gravitational lens aligned directly between a viewer and a more distant light-emitting object.

According to Einstein, who in the 1930's proposed how gravitational lensing could occur, perfect alignment should cause light passing at an equal distance from all sides of the lens to bend inward and converge on the viewer. If the lens is slightly out of alignment, the ring should break into two arcs; greater misalignment causes greater fracturing of the ring, until only a few small fragments appear. The only examples of gravitational lensing observed previously, which involve quasars split into two or three pinpoint images, are produced by badly aligned lenses, Chaffee says.

Petrosian recalls being disturbed initially that only one arc is visible in Abell 370 and 2244-02, but he and a graduate student, Anton Bergman, have created a model that shows how a lens consisting of two or more massive objects might create a single arc. Petrosian thinks two giant elliptical galaxies clearly visible inside the Abell 370 arc are the primary lensing masses. He estimates that in order to bend light over such a wide angle these giant galaxies must have at least 10 times more mass than is apparent from their luminosity. This estimate is in "rough accord," Lynds notes, with other estimates about the so-called dark matter that may be present in galaxies. *—J.H.*

In from the Cold

The new superconductors begin the long trip to market

Corporate and university researchers working with high-temperature ceramic superconductors have begun the race to carry the materials from the laboratory to the marketplace; it is likely to be a marathon, perhaps even an ultramarathon, with some jostling at the turns.

At Cornell University, Robert A. Buhrman says he has found a way to deposit thin superconducting films on a flat substrate to replace metals as the conductors that connect components of microelectronic circuits. The technique, known as high-pressure reactive evaporation, involves placing yttrium and barium inside a chamber held at 700 degrees Celsius and bombarding the metals with electron beams in an atmosphere of high-pressure oxygen. Buhrman and his colleagues report in *Applied Physics Letters* that the resulting metal vapors condense, along with metal vapor from a heated copper wire, on a substrate (typically zirconium oxide) and form a thin superconducting film of yttrium-barium copper oxide.

Buhrman maintains that the new method is a significant advance over a technique announced by the IBM Corporation last May: it eliminates a step and the process is carried out at lower temperatures that are less likely to damage delicate semiconductor components.

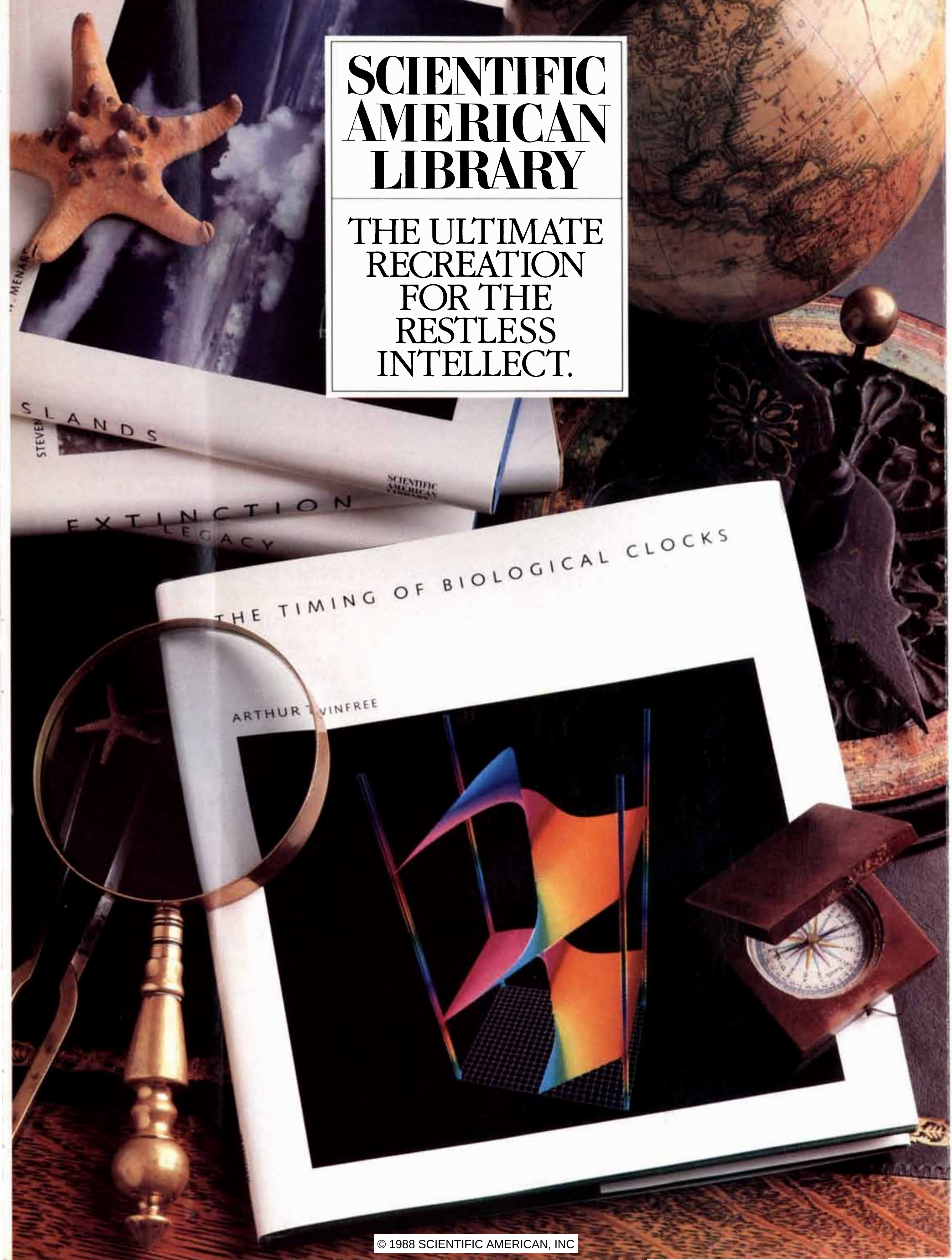
IBM has not been lounging in the blocks. Efforts are under way to deposit thin films of superconducting ceramic material directly on the silicon chips where the films are to link components. Unfortunately silicon destroys the film's superconductivity. To overcome this problem the workers, at IBM's Thomas J. Watson Research Center in Yorktown Heights, N.Y., are investigating various diffusion barriers that might be laid down over the silicon.

The other focus of competition is the preparation of bulk superconductors so that coils of them can carry enough current at high magnetic fields for such advanced applications as magnetic imaging equipment, magnetically levitated trains and the magnets of supercolliders. At the AT&T Bell Laboratories workers have developed a method that enables bulk superconductors to carry 100 times as much current as conventionally sintered superconductors.

According to Gilbert Y. Chin, a metallurgist at Bell Laboratories' Murray Hill, N.J., facility, the new method, known as melt-textured growth, involves melting oxide powders and then cooling the liquid. The resulting needlelike crystals, he says, conduct better than the smaller, more spherical crystals produced by sintering. Moreover, there are fewer grain boundaries between crystals, and he thinks that may eliminate the "weak links" in conduction. Chin says he is optimistic that conduction can be boosted by a factor of from 10 to 100 to commercial-application levels.

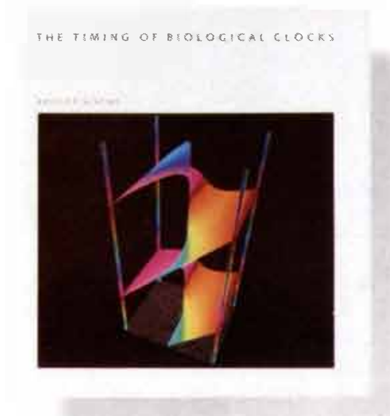
In the meantime tantalizing statements—none yet corroborated—continue to be made about superconductivity at room temperature. The latest is a preliminary report from Ahmet Erbil of the Georgia Institute of Technology of a solid that shows signs of superconductivity at temperatures as high as 500 degrees Kelvin (227 degrees Celsius). The material contains the same yttrium, barium, copper and oxygen as earlier authenticated superconductors but in slightly different proportions.

As for applications, the advice is not to hold your breath. "There is a large difference between synthesizing such materials," says John K. Hulm, chief scientist at the Westinghouse Research Laboratories, "and manufacturing them. If you want to use high-temperature superconductors in a superconducting supercollider, you will have to wait 10 or 15 years before you begin to build the machine." *—Gordon Graff*


SCIENTIFIC AMERICAN LIBRARY
THE ULTIMATE RECREATION FOR THE RESTLESS INTELLECT.
EXTINCTION
THE TIMING OF BIOLOGICAL CLOCKS

THE TIMING OF BIOLOGICAL CLOCKS

Arthur T. Winfree

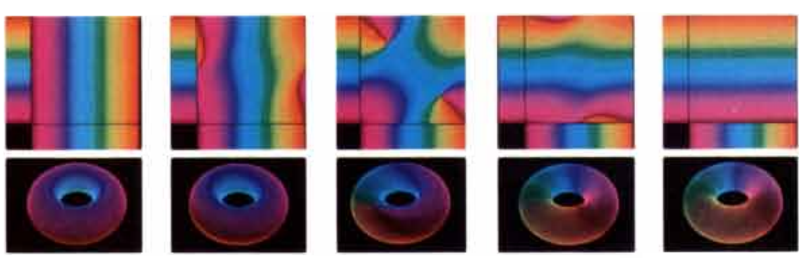


In 1931, Wiley Post flew around the world in eight days and became the first human being to experience jet lag. The fatigue and disorientation he felt happens to all of us when rapid long-distance travel knocks our "internal clocks"—our

In the early 1930s, Wiley Post used the Winnie Mae to study jet lag and its effect on pilot performance. Courtesy of the National Air & Space Museum, Smithsonian Institution.

circadian rhythms—out of kilter with local time. Arthur Winfree describes jet lag as "that disconcerting sensation of time travelers that their organs are strewn across a dozen time zones while their empty skins forge boldly into the future."

Jet lag, biorhythms, mosquito insomnia, temporal isolation experiments, the sleep movement of plants, "forbidden phases" of sleep when one cannot awake spontaneously—these are some of the many fascinating aspects of circadian rhythms that Dr. Winfree explores in *The Timing of Biological Clocks.*



Winfree's innovative use of gradient color to express the passage of time helps us visualize the biological cycles that govern the processes of life. Courtesy of Arthur Winfree.

Winfree shows that the most critical property of biological clocks is their ability to be reset on cue, enabling them to regain synchrony with a changing environment (as when we travel across

Flowering in the morning glory, as in many other plants, is timed by a circadian clock. Courtesy of Travis Amos.

time zones) or to adjust the body's 25-hour rhythm to the 24-hour solar day.

Reporting experiments on animals, plants, and single cells, he not only illustrates the principles that guide the resetting of biological clocks but reveals that each of these clocks has a vulnerable phase when a suitably intense cueing stimulus can produce a thoroughly unpredictable resetting—perhaps even annihilating the clock's rhythm entirely.

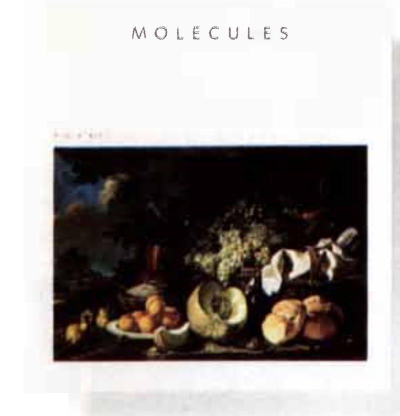
The graphics that Winfree uses are as innovative as his insights. By using gradient color rather than the conventional clock dial to express the passage of time, Winfree helps us visualize the true continuities—and discontinuities—of the internal cycles that govern the processes of life.

Has this cave salamander, living in temporal isolation, lost its circadian rhythmicity? Courtesy of Chip Clark.

Arthur T. Winfree is one of the world's foremost theoreticians of circadian rhythms. Trained as a biophysicist, Winfree received a MacArthur Grant for his work on biological clocks. Formerly at the Institute for Nonlinear Science at the University of California at San Diego, he is now with the Department of Ecology and Evolution at the University of Arizona at Tucson. Professor Winfree is also the author of the classic work *The Geometry of Biological Time* and *When Time Breaks Down,* a technical monograph on circadian rhythms.

199 PAGES, 233 ILLUSTRATIONS

P. W. Atkins



Everything we breathe, touch, eat, drink, and smell is made of molecules. They determine the taste of a barbecue and the fragrance of a rose. Yet, to many of us, molecules are an unfamiliar, abstract notion, hidden far below the surface of visible reality.

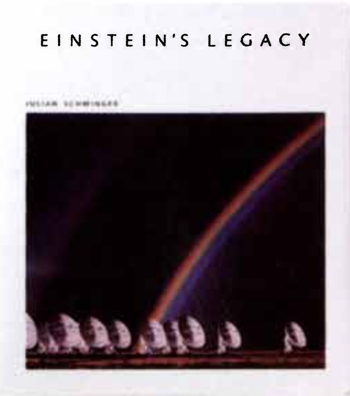
In this highly readable work, P. W. Atkins now makes the molecular world understandable, accessible, and familiar. You will meet the molecules that make sugar sweet, that make soap "soapy," and that give rise to the brilliant palette of colors unveiled in autum foliage. You will see molecules that can quell pain and alleviate depression as well as those that can induce addiction and wreak havoc on our environment.

Lavishly illustrated and filled with original insights, MOLECULES portrays the structures of the substances that make up our everyday world and our daily experiences, showing us their shapes, their interrelations, why they behave as they do, and how they carry out their tasks. It gives us a more imaginative appreciation of the richness, economy, and variety of nature.

P. W. Atkins lectures in physical chemistry at the University of Oxford. He is the author of *The Second Law* (for the Scientific American Library) and the widely used textbook *Physical Chemistry,* now in its third edition.

256 PAGES, 290 ILLUSTRATIONS

Julian Schwinger



Nobel laureate Julian Schwinger tells the captivating story behind one of the twentieth century's greatest achievements, the theory of relativity.

The groundwork was laid first by Isaac Newton and then by James Clerk Maxwell. But it was left to the sixteen-year-old Einstein to glimpse, for the first time, that the unifying theories of these two giants were incompatible.

In a lively and compelling narrative, Schwinger recounts Einstein's quest for the reconciliation of the conflict—a quest that ultimately led to the unification of matter and energy and of space and time in Einstein's special and general theories of relativity.

Schwinger carries this story to the space-age experiments that have been designed to confirm Einstein's legacy—and possibly lead to a still more comprehensive theory.

Julian Schwinger was awarded the Einstein Prize in 1951, the National Medal of Science in 1964, and the Nobel Prize for Physics in 1965. He is currently University Professor of the University of California.

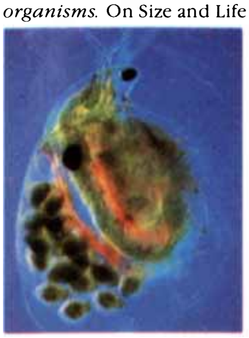
"Delightful and instructive.... An ideal gift for the curious non-expert."—NATURE

250 PAGES, 200 ILLUSTRATIONS

"The difference [between this and other book series] comes in the quality of the c

"Distinguished researchers take the trouble to talk simply and directly about the work they a

The largest forms of life reproduce at a slower rate and live much longer than smaller organisms. On Size and Life *examines why.*



A molecular view of the plasma membrane. Electr axopod of the heliozoon Echinosphaerium nucleof

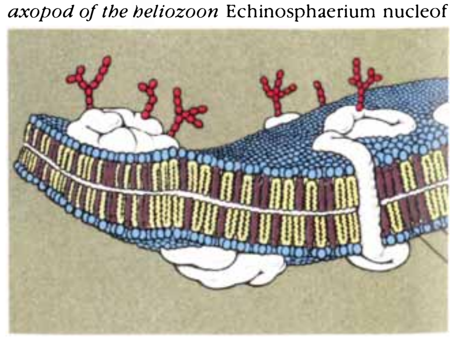
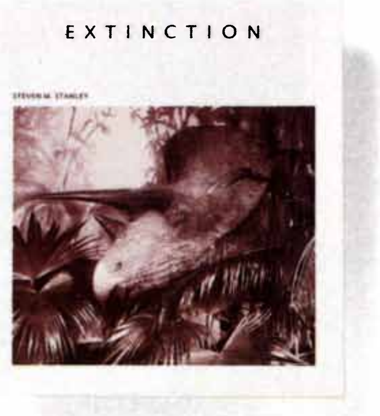


PHOTO CREDITS—Daphnia: James Bell, Photo Researchers, Inc./Hummingbird: Crawford H. Greenewalt, VIREO/Robert Wadlow, 8′ 9 1/2″ tall, and his father: Wide World Photos/Elephant: Peter Jo Black Star/Important anatomical divisions of the brain: Painting by Bill Andrews/Coca shrub, source of cocaine: W. H. Hodge, Peter Arnold, Inc.

Steven M. Stanley



Since the rise of multicellular life, a handful of mysterious cataclysms has swept our planet. These geologically brief outbreaks of mass extinction have decimated tens of thousands of thriving species, from huge dinosaurs to microscopic algae.

In EXTINCTION, Steven M. Stanley gives us a fascinating account of these dramatic global disasters, investigating what the rich storehouse of rock and fossil clues—and the revolutionary findings of plate tectonics—can tell us about their origins, progress, and implications.

In light of recent speculation that meteor impacts may have ended the reign of the dinosaurs through an early version of "nuclear winter"—about which Stanley has much to say—EXTINCTION is a timely as well as intriguing geological detective story.

Steven M. Stanley, professor of paleobiology and director of graduate studies at Johns Hopkins University, is a Guggenheim Fellow and winner of the Schuchert Award of the Paleontological Society.

242 PAGES, 183 ILLUSTRATIONS

OTHER VOLUMES IN THE LIBRARY:

MATHEMATICS AND OPTIMAL FORM
Stefan Hildebrandt and Anthony J. Tromba

ISLANDS
H. W. Menard

PERCEPTION
Irvin Rock

DRUGS AND THE BRAIN
Solomon H. Snyder, M.D.

THE SECOND LAW
P. W. Atkins

CONSTRUCTING THE UNIVERSE
David Layzer

THE LIVING CELL
Christian de Duve

POWERS OF TEN
Philip and Phylis Morrison and the Office of Charles and Ray Eames

FIRE
John W. Lyons

ON SIZE AND LIFE
Thomas A. McMahon and John Tyler Bonner

SUN AND EARTH
Herbert Friedman

EYE, BRAIN, AND VISION
David H. Hubel (in preparation)

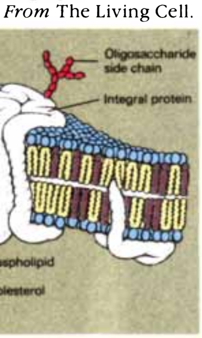
THE SCIENCE OF STRENGTH
J. E. Gordon (in preparation)

SAND
Raymond Siever (in preparation)

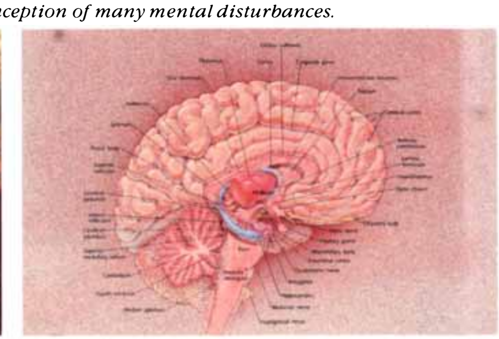
…ors, who combine eminence in their fields with proven gifts of exposition."
—Nigel Calder, NATURE

…doing and how it is solving yet another part of the mystery of the human and natural worlds."
—Steven Rose, NEW SCIENTIST

…crograph of a section through an … From The Living Cell.



Drugs and the Brain *explores the drugs that have led to a new understanding of the brain and that have changed our conception of many mental disturbances.*



Natural History Photographic Agency/Plasma membrane illustration by Neil O. Hardy/Electron micrograph courtesy of Dr. L. E. Roth and Dr. Y Shigenaka/Timothy Leary: Gene Anthony,

Off and Running

The Cheetah Masterplan may be insurance against extinction

The cheetah, the only mammal in the world that can sprint at speeds greater than 70 miles per hour, is on a fast track to extinction. Two factors threaten its existence. One is lack of genetic variation, which manifests itself in reproductive problems, excessive infant mortality and vulnerability to disease. These seem to be exacerbated in captivity, where cheetahs are experiencing a precipitous decline in number. Of 200 individuals in U.S. zoos, only 12 are actively reproducing, a level well below that necessary to sustain the captive population.

The other threat to the cheetah is loss of its natural habitat to agricultural expansion. In Namibia and other regions of Africa farming has steadily preempted the open range on which cheetahs hunt gazelle, antelope and other prey. Farmers, angered by the loss of livestock to these predatory cats, have won the right to legally trap and shoot them. Government records indicate that 700 cheetahs were killed this way in 1985; as few as 2,000 may be left.

Laurie Marker of Wildlife Safari in Oregon and International Studbook Keeper for the cheetah (an appointment made by the International Union of Directors of Zoological Gardens for the purpose of coordinating reproduction among captive cheetahs) has devised an innovative plan to thwart extinction. The Cheetah Masterplan, endorsed by the American Association of Zoological Parks and Aquariums, hopes to ensure survival of the animal through the creation of a viable captive population.

The first step calls for a rigorous analysis of such factors as diet, stress and behavior, which are believed to affect fertility and mortality among cheetahs. In this way it is hoped the reproductive problems plaguing them in captivity can be overcome and a self-sustaining population created. Eventually animals bred in captivity may be returned to their natural habitat to bolster declining numbers in the wild.

In order to begin the project its coordinators need to infuse the existing zoo population with a fresh group of reproductively healthy animals. Enter the Namibian farmers. Instead of exterminating cheetahs on their land, farmers there have agreed to give trapped animals to government officials for possible shipment to the U.S. Each animal being considered for the captive breeding program will undergo a comprehensive veterinary examination and only those that are reproductively healthy will be exported to the U.S., where they will be distributed on arrival among participating zoos. The plan has been received enthusiastically by both U.S. and Namibian officials and, if all goes well, the first 50 cheetahs will be shipped this summer.

"Our goal," says Stephen J. O'Brien of the National Cancer Institute, who is a principal adviser to the Cheetah Masterplan, "is to persuade farmers to trap the animals instead of shooting them. We have no intention of carrying out this project at the expense of the natural population." —*Laurie Burnham*

Darwin in Miniature

A monkey enzyme may show natural selection's handiwork

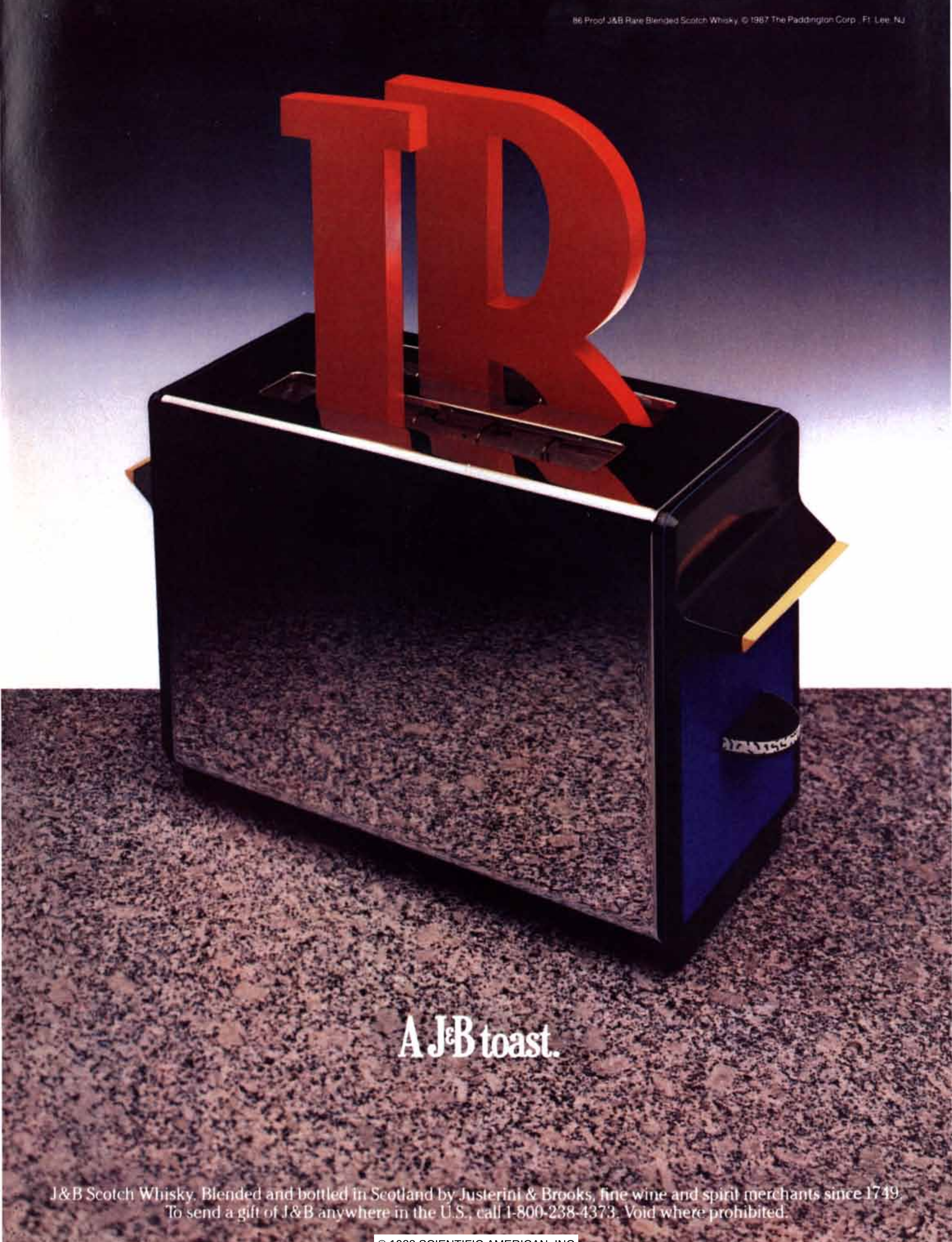
It is easy to see how natural selection could have produced an opposable thumb or a thick coat of fur. It has proved much harder to demonstrate the effects of Darwinian selection on a finer scale: to show that the specific chemical sequence of a protein or a gene represents an evolutionary adaptation to an organism's environment or way of life. The digestive specializations of the langur monkey have now enabled Caro-Beth Stewart and Allan C. Wilson of the University of California at Berkeley and James W. Schilling of California Biotechnology, Inc., to present an elegant description of adaptive change in a molecule.

The workers faced the challenge of disentangling adaptive evolution from the many molecular changes that seem to result from "neutral drift." Comparing the sequence of a specific protein or gene in various species generally reveals about the same rate of change in each species and each interval of geologic time—a pattern quite different from the one expected if the changes reflect selective pressures. To many investigators the constancy indicates that most sequence changes must be neutral: they neither help nor hinder the organism and become established only by chance.

Writing in *Nature,* the California group reports that in langur monkeys an enzyme called lysozyme has evolved in a manner decidedly inconsistent with neutral drift. In most mammals lysozyme is found in tears, saliva and white blood cells, where it serves to fight invading bacteria by breaking down the bacterial cell wall. Ruminants such as cows and sheep have put lysozyme to another use in coping with their high-cellulose vegetarian diet. Their specialized "foregut" serves as a fermentation chamber, in which bacteria break down cellulose; in the stomach proper the bacteria themselves are digested with the help of lysozyme. The same digestive specializations arose independently in the leaf-eating colobine monkeys, among them the langur.

The workers determined the amino acid sequence of langur lysozyme and compared it with the enzyme's sequence in baboons, human beings and several other animals, including cows. Evolving solely by neutral drift, the lysozyme of each animal should have diverged at a steady rate from the lysozyme of every other animal. In the langur and the baboon, for example, the enzyme should have changed by an equal amount since they last shared an ancestor with humans; all three primate sequences, in turn, should have diverged equally from the cow sequence. Instead comparisons with the human sequence suggested that langur lysozyme has evolved twice as fast as the baboon enzyme. What is more important, at about half of its altered amino acid positions langur lysozyme has gained similarity to lysozyme from the cow.

Together, the investigators write, the results could imply that "positive Darwinian selection has driven about 50 percent of the evolution of langur stomach lysozyme." The selective pressures presumably arose as lysozyme assumed a digestive role in langurs analogous to the role it plays in cows. Stewart and Wilson have also studied the significance of the convergent sequence changes. In the 1987 *Cold Spring Harbor Symposium* they suggest that the specific amino acids shared by langur and cow lysozymes may help the enzyme to function in acidic stomach fluid and to resist attack by fermentation products and by digestive enzymes. —*Tim Appenzeller*


86 Proof J&B Rare Blended Scotch Whisky. © 1987 The Paddington Corp., Ft. Lee, NJ
A J&B toast.
J&B Scotch Whisky. Blended and bottled in Scotland by Justerini & Brooks, fine wine and spirit merchants since 1749.
To send a gift of J&B anywhere in the U.S., call 1-800-238-4373. Void where prohibited.

Fractal Shorthand

"Rule-based" algorithms let computers store more images

Fractals, geometric forms whose irregular details recur at different scales, are often fantastically complex. And yet one can create a fractal merely by plotting points on a sheet of paper according to two simple rules, randomly applied. One begins at any point. The rules then might stipulate that the next point plotted always be either halfway toward the upper right-hand corner of the page or one-third of the way toward the center and rotated clockwise 40 degrees; a coin flip decides which rule is applied. The points seem randomly distributed at first, but after hundreds of coin flips a distinct form—a fractal—emerges.

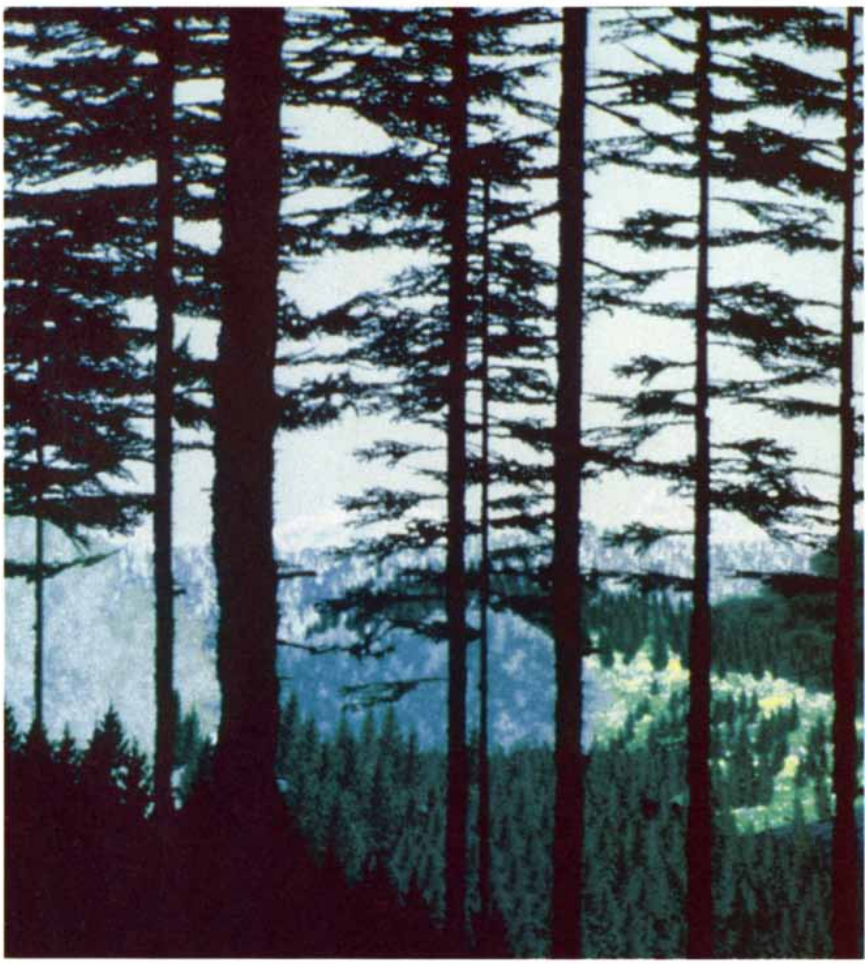
With enough rules and a computer one can generate not only fractal curiosities but also images of natural phenomena, according to Michael F. Barnsley and Alan D. Sloan of the Georgia Institute of Technology. The workers have written rule-based algorithms that produce images of pine forests, the wakes of submerged submarines, human faces and many other objects. The chief benefit of their technique, they say, is that a complex image can be represented in electronic shorthand, not in its full digital form but in terms of its rules.

With 2,000 bytes of memory, according to Barnsley, a computer can store the rules needed to create a million-pixel color picture of a pine forest; normally a computer would require as many as three million bytes to store such an image. This image compression, Barnsley says, allows more images to be stored in the memory of a computer and transmitted over cables or radio links.

Finding the right rules, Barnsley acknowledges, can be tricky. The process has been largely automated for black-and-white images of objects that are obviously fractal-like, such as clouds, ferns and mountains. Finding rules for more complex images, such as a color picture of a human face, is still difficult and time-consuming. An investigator must try to break the picture into components that have fractal qualities and then devise a separate set of rules for each component. After further research and the development of specialized hardware, Barnsley says, the rule-writing process may become fully automated for all types of images.

Barnsley thinks the research could benefit what he calls the "classical computer-graphics industries," such as computer-aided design, flight simulation and cinematic special effects. The U.S. Defense Advanced Research Projects Agency, which is funding the research, is more interested in automatic target recognition. The agency hopes, for example, to improve the ability of satellites to detect the surface turbulence above submerged submarines; modern submarines, Barnsley notes, create much less turbulence than their predecessors. Fractal algorithms could speed a satellite's search for a match between what it sees and what is in its memory. The satellite could also transmit the compressed images faster to analysts on the ground. *—J.H.*

The fractal essence of forests, faces and the wakes of submarines can be expressed in simple rules



FRACTAL FOREST, based on a photograph of Germany's Black Forest, was produced by Michael F. Barnsley, Arnaud E. Jacquin, François J. Malassenet, Laurie B. Reuter and Alan D. Sloan of the Georgia Institute of Technology Research Corporation.

Wave Watch

A new tsunami-warning system is fast, reliable and cheap

Tidal waves triggered by undersea earthquakes or volcanic eruptions have killed some 52,000 people around the Pacific basin in the past 100 years. Systems that warn populations of an approaching wave must work quickly: most deaths caused by these waves, called tsunamis, occur in the first 20 minutes and less than 100 kilometers from the source. Developing countries cannot afford

Nikon introduces the p

You don't have to be a genius to use the new Nikon N4004.

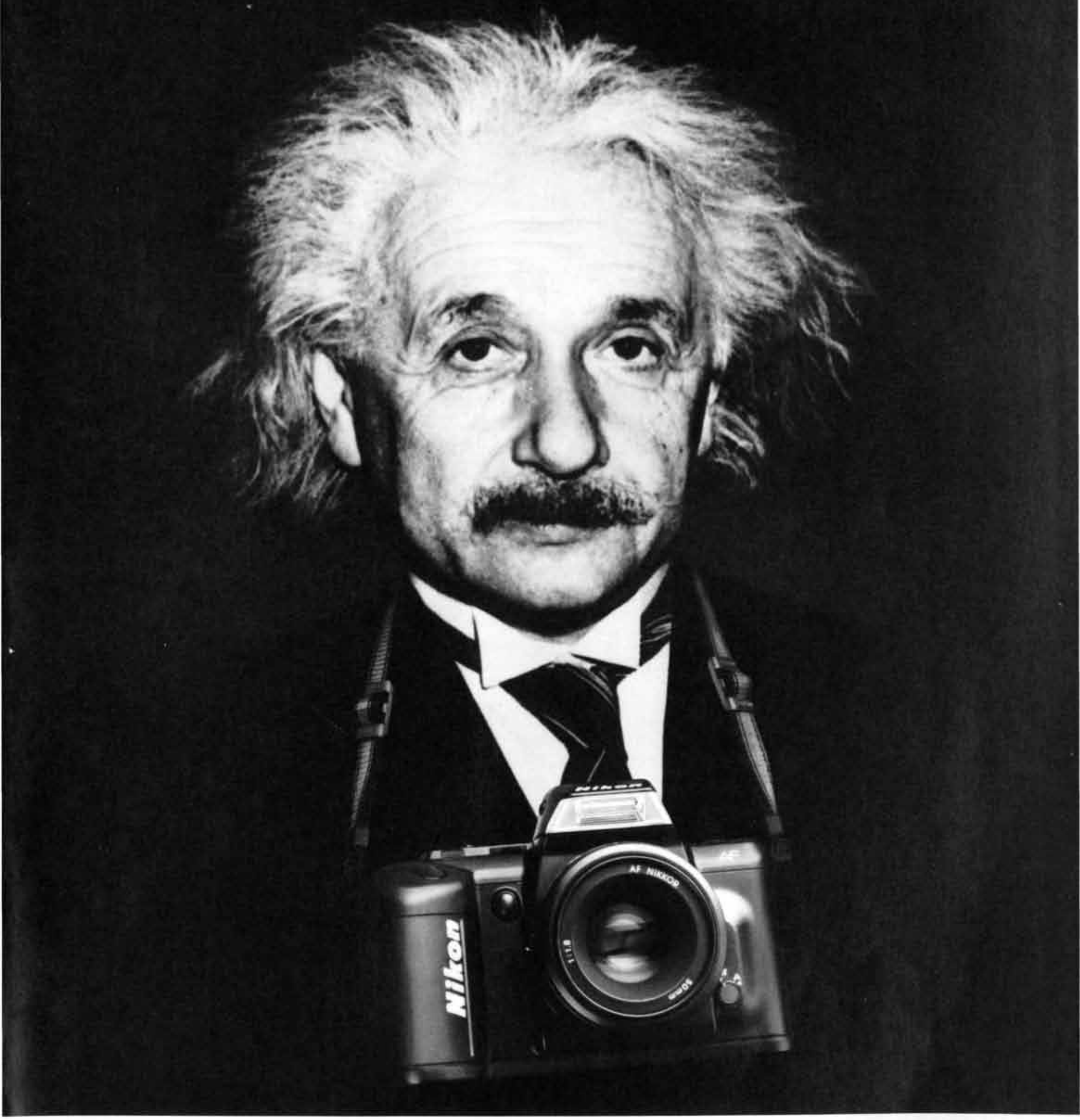
Even if your photographic IQ is near zero, this is one 35mm SLR you can take out of the box and begin using right away. Because it does everything for you.

It incorporates a remarkable Nikon innovation called the Decision Master System, which controls all camera, lens and flash functions automatically, even in difficult lighting situations.

The N4004 also loads, advances and rewinds the film automatically. It even focuses automatically.

One year membership in the Nikon USA Club is free with every purchase when you submit the Nikon USA Club application. For further information write: Dept. N15, Nikon Inc., 19601 Hamilton Ave., Torrance, CA 90502-1309. © Nikon Inc., 1987.

erfect camera for both.



When you need a flash, the N4004 will recommend that you use one. And you'll always have one, since the flash is built in.

But most important of all, as your photographic genius grows, and you want more creative freedom, the N4004 becomes less automatic. Allowing you to make all the settings yourself.

The Nikon N4004 is incredibly easy-to-use. At the same time, it's incredibly sophisticated.

That's not a contradiction. That's genius.

Nikon. We take the world's greatest pictures.®

Moe Howard © 1987 Norman Maurer Productions, Inc. Columbia Pictures Industries, Inc. Einstein licensed by The Roger Richman Agency, Inc., Beverly Hills, CA.

the sophisticated systems—costing upward of $1 million to install and $500,000 per year to operate—that serve Hawaii, French Polynesia, Alaska, Japan and the Pacific coast of the Soviet Union.

But now the developing countries have a cheap, simple alternative. Using off-the-shelf instrumentation and one of its own satellites, the U.S. National Oceanic and Atmospheric Administration (NOAA) has developed a system that costs $20,000 to install, can be run by nonexperts and transmits a warning in a single minute—10 times faster than the big systems do.

Sponsored by the Agency for International Development, an NOAA team built a pilot system in Valparaiso, Chile, a Pacific coast city where 18 tsunamis have struck in the past 200 years. Because of its simple, modular design, the pilot stands as a "blueprint of an effective systems approach to hazard mitigation," according to Eddie N. Bernard, leader of the team.

Working with Chilean officials, the NOAA team used historical data and computer simulation to develop an evacuation plan for the sections of the city most in danger from rapidly rising water. The team then installed a commonly used sensor called an accelerometer in the bedrock under Valparaiso and water-level sensors in the city's harbor. When an earthquake measuring 7.0 or more on the Richter scale occurs within 60 miles of the coast, the accelerometer senses the event and transmits a signal to an NOAA weather satellite. The satellite beams an alert code to a computerized receiver in Valparaiso. The computer is programmed to send emergency messages over telephone lines to various officials; it also automatically begins monitoring the water-level sensors.

This instrumentation, Bernard observes, is durable, battery-powered and affordable (the pilot system employed a Commodore 64). Each component can be upgraded or serviced separately, and the drain on the satellite is negligible.

No tsunamis have occurred since the pilot system was installed. Still, in 5,000 random tests during the past year, it proved to be 98 percent reliable in recognizing earthquakes over the threshold size. The more expensive and elaborate systems, on the other hand, which provide continuous, precise monitoring of seismic activity and water levels with state-of-the-art sensors, are only about 90 percent reliable and take from 10 to

15 minutes to transmit a warning, according to Bernard. He maintains that sacrificing precision for speed makes sense; after all, during a rapid evacuation, he points out, "who cares whether the earthquake is 7.0 or 7.8?" *—Elizabeth Collins*

Ritual on Wheels

Or why the Mesoamericans never got on a roll

When most people think of ancient Mesoamerica (Mexico and central America), they think of it without one of the basic components of European civilization: the wheel. They are right—but not completely. For some time it has been known that the ancient inhabitants of Mexico, El Salvador and Guatemala had small toylike wheeled devices that were never put to practical use. The discovery of the single largest sample of the wheeled figurines throws new light on the origin, spread and function of the wheel in Mesoamerica and on why the innovation was never exploited for transport.

The find was made at a site in central Mexico called Tula. Tula, founded in about A.D. 750, was the capital of Toltec society. (The Toltecs preceded the Aztecs, who looked on them as heroic ancestors.) Tula probably reached its height in about A.D. 1050, when it had some 35,000 inhabitants; not long after that Toltec culture collapsed.

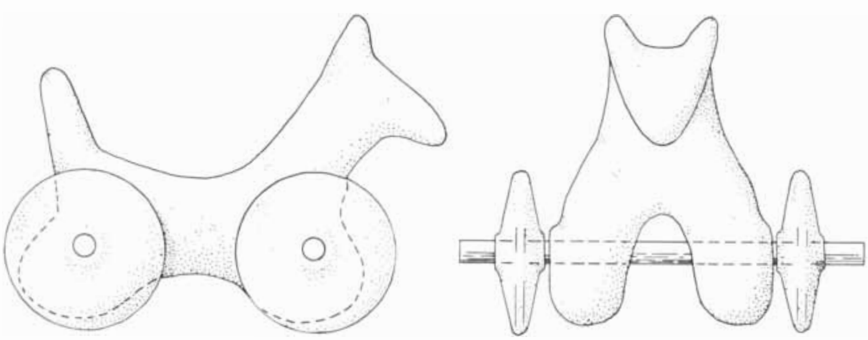
A team led by Richard A. Diehl, who is now at the University of Alabama, carried out extensive excavations at Tula. In the northern part of the site an area that included 15 houses was dug; it was there that the fragments of the wheeled objects were found. The 79 fragments that were unearthed were the remains of at least 27 different figurines. The intact figures would typically have been about 10 centimeters long. Each wheeled effigy represented an animal with axle holes in each leg. Most of the heads depict dogs, animals having potent iconographic significance in ancient Mesoamerica.

What new insights does the find at Tula provide? First, the quantity of wheeled figures from there is much greater than has been reported elsewhere. "Previously there was one wheeled figure reported here, one there," Diehl said. "At Tula, however, the wheeled effigies represent almost 10 percent of the total of clay figurines. It suggests that these things may have been much commoner than we thought."

In addition, the location of the fragments suggests that the wheeled effigies may have been used in household ceremonies. Early investigators called the wheeled figurines toys. Recent work, however, suggests they were ritual objects, and Diehl concurs. The association of the wheeled Tula fragments with the ruins of houses rather than temples indicates that they were not employed in rituals carried out in the temple by a priestly elite. And what might the humble household rituals have been? Diehl admits to being stumped.

He adds that the evidence suggests that the concept of the wheel probably originated on the Gulf of Mexico in what is now the state of Veracruz not long before A.D. 900. Soon afterward it reached central Mexico and

The finding of a large number of clay figurines elucidates how ancient Mexicans used the wheel



WHEELED EFFIGY was reconstructed from fragments discovered at Tula, the ancient Toltec capital in central Mexico. About 10 centimeters long, the figurine depicts a dog.

El Salvador. Although the wheel was widespread, two factors explain why it never became practical, Diehl says. The most crucial is the lack of a suitable draft animal: the largest domestic animals available to the ancient Mexicans were medium-size hairless dogs called *Xoloitzcuintli.*

Another factor was Mexico's terrain. Both the central highlands and the swampy coastal lowlands of Mexico present serious obstacles to wheeled transport. Rather than exploiting the wheel, the ancient Mesoamericans developed a system of transportation based on backpacks and boats. Diehl concludes, as other workers have, that until the arrival of the internal-combustion engine the costs of using the wheel for practical purposes in Mexico—such as building roads and feeding large animals—considerably outweighed all its potential benefits. *—John Benditt*

MEDICINE

Exonerating Sodium

Table salt, not sodium itself, is implicated in hypertension

Has sodium been getting a bum rap? For a decade health organizations, government agencies and physicians have urged everyone to cut back on the intake of all forms of sodium.

R. Curtis Morris, Jr., of the University of California at San Francisco's General Clinical Research Center thinks such recommendations may be wrong. "It has never been demonstrated," he says, "that any form of sodium other than sodium chloride [table salt] can raise blood pressure." This means, for example, that the sodium ingested as sodium bicarbonate in baked goods or antacids probably does not affect blood pressure and should not be lumped together on food labels with the sodium of sodium chloride. Even the latter has not been shown to raise blood pressure in everyone, Morris says: only about half of the 40 million Americans who suffer from hypertension seem to be sensitive to table salt.

Morris bases his challenge to conventional wisdom on a study, published in the *New England Journal of Medicine,* that he did with Theodore W. Kurtz and Hamoudi A. Al-Bander. The investigators compared the effects of table salt and sodium citrate on five salt-sensitive men who had the commonest form of high blood pressure, essential hypertension. The subjects were hospitalized for six weeks and put on identical low-salt diets. Within a week their blood pressure dropped to levels that were close to normal. The men were then given capsule supplements containing sodium citrate, salt or a placebo in randomized weekly rotations. Each subject's blood pressure rose when he was given the table salt supplement; it dropped following a switch to sodium citrate or the placebo. Other investigators, Kurtz says, have shown that sodium compounds such as monosodium glutamate and sodium phosphate (a common ingredient in soft drinks) also appear not to affect hypertension.

Morris cautions that because of the expense of hospitalization the study group was small and so the data should be considered preliminary. Still, he says, the study has dietary implications. For people who consume large amounts of baked goods and processed foods, dietary restrictions based on sodium labeling may be unnecessarily harsh.

What about sodium chloride, then? Should everyone cut down on table salt? Both Kurtz and Morris respond negatively. Kurtz notes that many factors other than salt consumption can contribute to the development of hypertension. Furthermore, restricted salt intake may correlate with low blood pressure in one patient and high blood pressure in another. Morris asserts that no large-scale studies have proved that the long-term consequences of a high salt intake are harmful. In spite of such considerations, strictures about salt are likely to persist. "It is hard," Morris points out, "to be evangelical about moderation." *—E.C.*

New Genes for Old

A gene-therapy test may soon be proposed, but doubts persist

Investigators at the National Institutes of Health plan to ask within months for permission to treat a fatal genetic illness by providing a patient with a new working gene. The workers discussed the project informally with a regulatory committee in December so that when the technique is ready for testing and a prospective patient is available prolonged deliberations can be avoided.

The gene therapy would initially be used as a treatment of last resort for Severe Combined Immune Deficiency disease (SCID), which affects fewer than 40 babies a year in the U.S. If it is successful, the intervention could lead to similar therapy for other inherited diseases.

In the type of SCID that would be treated the patient carries a gene that produces a defective form of the enzyme adenosine deaminase. Failure of the enzyme thwarts the development of the immune system; untreated, the disease usually kills within two years. W. French Anderson and his colleagues at the NIH plan to remove bone-marrow cells from an SCID patient and infect them with a modified retrovirus carrying a human adenosine deaminase gene. The retrovirus is modified by genetic engineering so that it cannot replicate by itself; each virus should therefore invade just one bone-marrow cell and spread no farther. The marrow cells would then be returned to the patient, and the investigators hope that a functioning DNA copy of the deaminase gene will be stably incorporated into some of the cells, producing working enzyme.

Hope is the operative word. Although a similar procedure works well in mice, in treated monkeys fewer than one in 100 marrow cells were infected, and in only two out of five monkeys was there evidence that human gene activity persisted for a long period. Critics think this may mean that only the current generation of marrow cells were infected, rather than the stem cells from which they are derived.

Safety is also an issue: although no ill effects have been seen in experimental animals, the transfer virus contains sequences that promote gene expression and so could conceivably activate normally benign oncogenes present in all cells, thereby giving rise to cancer. A more serious concern is that about one in a 1,000 transfer viruses is able to replicate; this could in principle cause an infection and make the patient shed a virus that might have dangerous properties.

Such concerns led the regulatory committee to respond cautiously to an informal description of the approach submitted by Anderson. The committee is a working group of the

Recombinant DNA Advisory Committee, known as the RAC, the most important of the bodies that must approve a specific proposal to administer the therapy. Richard C. Mulligan of the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology, a member of the working group, says "no evidence has yet been presented or is on the horizon to show that this would actually work."

Anderson hopes he will be able to allay such doubts. He plans to reduce further the ability of the transfer virus to replicate, using genetic constructs that Mulligan has offered. He is also doing experiments with human cell cultures to provide evidence that the virus does indeed infect stem cells. A formal proposal will be made in "a few months" if no ill effects emerge, according to Anderson.

Mulligan says it would be "scientifically foolhardy" to attempt gene therapy now; he does concede that treatment might be ethically justifiable in patients for whom there is no other hope. —*T.M.B.*

Supertests

DNA amplification could simplify prenatal diagnosis

Clinicians searching for signs of viral infection in a patient or the presence of genetic disease in a developing fetus face the same problem that confronts the seeker of a needle in a haystack. Viral genomes inserted into the host's chromosomes and aberrant genes that must be analyzed represent only a tiny fraction of the DNA in each cell. In order to identify such a subtle pathogenic presence the clinician has had to rely on radioactively labeled, matching strands of DNA that combine with the target DNA. Because the pathogenic DNA is so scarce, relatively large amounts of host DNA must be collected. To get the hundreds of cells required from a fetus requires invasive procedures such as amniocentesis or the more recently developed chorionic-villus sampling; both methods pose some risk to the fetus.

A recent innovation skirts these problems by enabling clinicians to greatly amplify the number of needles in the genetic haystack. The technique, which can be automated, relies on the polymerase-chain reaction, developed in the past 18 months by Randall K. Saiki, Henry A. Erlich and their co-workers at the Cetus Corporation. Investigators make two short "primer" strands of artificial DNA that bind at each end of the stretch of target DNA that is to be amplified (each primer binds to a different strand of the double helix). A form of the enzyme DNA polymerase is then added, which copies DNA in opposite directions on the two strands once it has both a primer sequence to get it started and a template strand to copy from.

The result is that the DNA between the two primers, and no other DNA, is copied continuously as each copy gives rise to new copies; consequently the amount of target DNA grows exponentially. A refinement uses a heat-stable polymerase that works faster and improves the accuracy of amplification. In principle a target DNA sequence from even a single cell could be analyzed, and a test that previously took weeks can now be done overnight, according to Joe W. Gray of the Lawrence Livermore National Laboratory—an important consideration in prenatal diagnosis.

Cetus is developing several infectious-disease tests based on the principle. Such a test could, for example, detect infection with the AIDS virus directly, without depending on radioactive probes, while avoiding the relatively frequent false-positive and false-negative results that antibody-based tests are subject to. Some experimental clinical application will begin in 1988.

The prospect that a diagnosis can be made from a few cells has intensified the optimism of workers searching for ways to identify the small numbers of fetal cells thought to be present in an expectant mother's blood. Although the blood of an unborn child remains largely separate from that of its mother, some fetal blood cells may cross the placental barrier during early pregnancy. If they could be detected reliably, DNA amplification might allow a standard maternal blood sample to provide enough fetal cells for testing at virtually no risk to the fetus.

Investigators are now experimenting with fluorescent antibody probes that, they hope, bind only to fetal cells, making them glow when illuminated so that they can be separated by machine. "The crucial thing," says Leonard A. Herzenberg of the Stanford University Medical Center, who pioneered the technique, is "the confirmation that we can find something only from fetal cells." No one yet has that confirmation, but Diana W. Bianchi of the Harvard Medical School is "more optimistic than ever" that antibodies she is using can identify immature fetal red blood cells. Susan J. Fisher of the University of California at San Francisco School of Medicine is searching for another type of fetal cell, one that invades uterine blood-vessel linings; Herzenberg applies molecular probes that bind directly to chromosomes rather than to the cell surface. —*T.M.B.*

OVERVIEW

Flying Blind

Testing an AIDS vaccine may be harder than inventing one

In the U.S., where AIDS has already claimed almost 30,000 lives, no fewer than a dozen commercial and government laboratories are devoted to research on vaccines. Two AIDS vaccines have already been approved by the Food and Drug Administration for clinical trials, and through those trials more than two dozen people have been vaccinated. In Zaire, where last year Daniel Zagury of the University of Paris carried out the first AIDS-vaccine tests in human beings, extensive studies involving several hundred vaccinees are about to begin. This progress seems extraordinary if one considers that the human immunodeficiency virus (HIV) was discovered just four years ago.

Yet the mood at a recent AIDS-vaccine conference in Washington was described by the researchers there as "appropriately grim." Efforts to produce a vaccine for AIDS are being crippled by the fact that there is really no way of telling whether the vaccine works short of giving it to humans. Only people get AIDS; chimpanzees can be infected with the virus, but so far none of the infected chimps have developed the disease. Without a proper animal model it is anyone's guess whether the imminent crop of AIDS vaccines will be effective.

There are certainly enough unknown angles to HIV infection to squash optimistic speculation. Variability among strains and HIV's high mutation rate make the virus an elusive target. No vaccine has ever been made against a human retrovirus—the genre of virus to which HIV belongs—and no one knows which of the immune system's defenses must be activated in order to stop the virus from causing disease.

Clinical trials may help to dispel some of these gray areas. Unfortunately there are many indications that clinical testing itself may be the most problematic aspect of AIDS-vaccine development. Safety is not an issue, because most vaccines contain only pieces of HIV and therefore pose no threat of causing the disease they are meant to prevent. But the ongoing tests are the first U.S. vaccine trials ever to be undertaken without prior evidence of efficacy.

The closest thing to evidence of how experimental vaccines might perform in humans comes from data on chimps. The results are ambiguous at best. Typically, the vaccinated chimps produce antibodies that knock out the virus in a test tube, but when they are challenged with injections of HIV, they still become infected. As the need for a model grows more urgent the scientific community has been scrambling for creative solutions. Some investigators think mice could be genetically engineered to contract AIDS. Others look to analogues of HIV, such as simian immunodeficiency virus or feline leukemia virus, to provide clues to mechanisms of immunity.

In the meantime applications for human testing of experimental vaccines are piling up at the FDA, forcing decisions that cannot wait for the ideal animal model. The issues are fairly straightforward for the first phase of clinical trials, which are designed only to assess at various dose levels a vaccine's safety and its immunogenicity: the extent to which it elicits an immune response. Whom should the vaccine be tested on first? What criteria should be established with respect to volunteers' sexual practices? How can volunteers be protected from the stigma of being HIV-seropositive—testing positive for antibodies to the AIDS virus?

The precedent for these decisions was set last August when a vaccine that was produced by MicroGeneSys, Inc., in West Haven, Conn., won the first approval for clinical trials. The trials, which are being conducted by the National Institute of Allergy and Infectious Diseases (NIAID), will involve only low-risk, HIV-seronegative men: 78 homosexuals and three heterosexuals. The volunteers are identified as "low risk" if they could not have been exposed to HIV in the preceding three months. They are instructed not to engage in any sexual behavior that could result in infection during the course of the trials, which will last from six months to a year. Although it is possible to distinguish the immunological profile of an infected individual from that of a vaccinated individual, the participants will be issued cards explaining the source of their HIV-seropositive status. Investigators at the University of Washington School of Medicine will follow a similar protocol in this year's trials of a vaccine made by Oncogen in Seattle.

Even at this early stage the NIAID has had some trouble recruiting volunteers. As of mid-December, after three months of nationwide advertising and several hundred inquiries, the institute had found fewer than 30 eligible participants. This scarcity is not too surprising: unlike participants in trials of therapeutic drugs or vaccines of demonstrable efficacy, the volunteer in an AIDS vaccine trial has everything to lose and nothing to gain. "It became clear to us very early on that asking healthy individuals to participate for reasons other than serving mankind or being Rambos didn't make ethical sense," says David T. Karzon, a professor of pediatrics at the Vanderbilt University School of Medicine who has worked extensively at the Federal vaccine-evaluation center there. "We don't know that there's any gain for the individual."

The situation could change if results from the first phase of trials are particularly encouraging. Only another 100 or so people will be needed for the second phase of testing, which will extend the immunogenicity and dose studies. It is not clear, however, what immunogenic results should be interpreted as "encouraging," because no one knows whether the production of antibodies or the activation of immune cells plays the more important role in protection from HIV. When the vaccines enter the third phase of testing, new choices will have to be made. Phase III trials raise the question of efficacy: Will this vaccine keep a person—man or woman, homosexual or heterosexual, adult or child, infected or uninfected—from getting AIDS?

A Phase III study usually involves thousands of vaccinees and takes several years to complete. How will the clinical establishment scrape up thousands of participants when it is having trouble finding 80? As more vaccines enter trials the reservoir of recruits will be depleted even further. Eventually tests will have to be extended to the heterosexual population, and there the demands of statistical significance will be particularly hard to meet because the incidence of infection and disease in that population is extremely low. The systematic screening of some cohorts, such as military recruits and the wives of hemophiliacs, has turned up subpopulations with unusually high rates of infection that could be used for efficacy studies.

The as yet undetermined latency period of HIV will also present a quandary for Phase III trials. If a vaccinated individual were to become infected but not develop the disease, it could be said that the vaccine provided protection. But unvaccinated individuals have been known to harbor latent HIV for years and then suddenly and inexplicably fall victim to AIDS. How long should researchers wait before they pronounce a vaccination a success? Should infection in itself be considered failure?

The U.S. Government has acknowledged the solemnity of the AIDS crisis with more generous support for vaccine research every year; the 1987 allocation of $31 million is more than 10 times what was awarded in 1984. This year the NIAID plans to fund between five and seven AIDS-vaccine development groups and to establish a repository for the cell cultures and viral isolates the workers need in their studies. These steps may facilitate the scientific understanding of HIV's attack on the immune system, but they do not address the need for an animal model or for a thorough appraisal of how to proceed with clinical trials.

MicroGeneSys' vaccine, the first to be approved, could be ready for efficacy tests in just two years. There will probably be other candidates fast on its heels. The general availability of an AIDS vaccine could be delayed unless by that time the scientific community has reached a consensus on the protocol for efficacy testing. Even optimistic estimates suggest that a vaccine will not be available before the mid-1990's, by which time millions of Americans could be carrying the virus. Delay could cost lives and haste could have its own damaging consequences, but in the absence of an animal model for AIDS, human beings will have to be the guinea pigs. —*Karen Wright*

That's a fairly odd phrase — "when you realize the future's behind you." But a forward thinking company understands and knows exactly what it means.

Quite simply, the kind of company you are tomorrow depends a great deal on the moves you made yesterday, the decisions you make today.

In recent years the moves we've made, the decisions we've reached have renewed our spirit. It took introspection, belt-tightening, and some tough calls. Now we're smarter, more flexible, and more responsive to change.

It all strengthened our belief that the way to step toward tomorrow is not to follow, but to lead. With innovative thinking that allows us to do business as we've never done it before.

Today we're involved in emerging technologies that can lead to new business opportunities tomorrow. Projects like the development of synthetic metals for increased safety in commercial aviation. A new venture with the Baylor College of Medicine to produce an early cancer detection kit. And a proprietary synthesis technology for the formulation of nontoxic pest control.

When you look back at where we've been, then forward to where we're going, you'll find we're a company recharged and marching on.

For more information please write to Patricia Marshall, Phillips Petroleum, 16D-4 Phillips Building, Bartlesville, Oklahoma 74004.

Phillips Petroleum Company
Performing to meet
the challenge of change.

PHILLIPS 66

WHEN YOU REALIZE THE FUTURE'S

SCIENTIFIC AMERICAN February 1988 Volume 258 Number 2

Chromosome Mapping with DNA Markers

Variable sequences in the DNA of human chromosomes act as genetic landmarks. Individual markers serve for tracing defective genes; collectively the markers provide the elements of a chromosome map

by Ray White and Jean-Marc Lalouel

Say that a disease is known to run in families, following a classic Mendelian pattern of inheritance. Somewhere among the 100,000 genes on the 23 pairs of human chromosomes a single gene is defective. The symptoms and progress of the disease have been described in meticulous detail, but its biochemistry is an enigma, and even predicting who will actually get the disease is guesswork. Such has been the case not just for a handful of rare afflictions but for most of the 3,000 known genetic diseases, including such familiar scourges as Huntington's disease and cystic fibrosis. Where does one begin the search for a causative mechanism, a diagnostic test and, ultimately, a treatment?

It is now possible to start by closing in on the defective gene itself. The territory to be surveyed is vast: the human chromosomes consist of linear molecules of double-strand DNA with a total length of about three billion base pairs (the chemical subunits that encode information along DNA). A typical gene, a complete unit of genetic information, is minuscule by contrast, encompassing perhaps 10,000 base pairs. And yet by correlating the inheritance of a distinctive segment of DNA—a "marker"—with the inheritance of a disease, one can now localize the mutant gene to within one or two million base pairs, or less than a thousandth of the human genome (the total complement of DNA). That kind of precision puts the gene within reach of molecular tools for cloning DNA and testing its activity. The identification of a genetic marker that is closely linked with a disease also means the gene's inheritance can be followed. It opens the way to simple tests for diagnosing carriers and future disease victims.

The basic strategy, known as linkage analysis, is a venerable tool of classical genetics. In our laboratory at the University of Utah and in many others, however, it has gained new power from the techniques of molecular biology, which make available a greatly expanded set of markers: molecular variations known as RFLP (for restriction-fragment length polymorphism) markers. Linkage analysis has now revealed RFLP markers for a number of disease genes, and many more diseases will soon yield to the strategy. It is also serving a more general purpose. By following the inheritance of many RFLP markers simultaneously in healthy families, we and other workers have begun to plot their positions in relation to one another and map them onto the physical framework of the chromosomes. The goal is a complete map of markers: an array of reference points that spans the genome and makes it possible to pinpoint disease genes far more efficiently than can be done with isolated markers.

The linkage strategy exploits the way genes are inherited. An ordinary human cell contains 23 pairs of homologous, or matching, chromosomes, one chromosome per pair inherited from the mother and the other from the father. In meiosis, the series of cell divisions that gives rise to germ cells (sperm or eggs), the homologous chromosomes in a progenitor cell are duplicated and then distributed among four germ cells, each of which receives 23 single chromosomes. The parental chromosomes are not transmitted intact, however. In the course of meiosis homologous chromosomes repeatedly recombine: they "cross over" and exchange segments of equal length [*see illustration on page 43*]. As a result each chromosome that is transmitted in a germ cell is generally a patchwork of segments from the two parental chromosomes. Recombination is the phenomenon that enables one to find linkage between a marker and a disease.

What makes it possible to detect recombination and employ it in linkage analysis are the many differences between homologous chromosomes. They often carry two different alleles, or versions, of many of their matching genes and also of many apparently meaningless DNA sequences within and between genes. The recombinant chromosomes that are parceled out to the germ cells at meiosis represent new combinations of these features. An allele from a locus on one chromosome and an allele from a different locus on the other, homologous chromosome can be combined and passed on together; at the same time the alleles at two loci on a single chromosome can be separated, so that only one of them is inherited.

The closer together two loci lie on the same parental chromosome, the less often their alleles are separated as DNA is exchanged between homologous chromosomes during meiosis. Hence one can gain a measure of the distance between a gene of particular interest—one that has a disease-causing mutant allele, for example—and a marker by correlating the inheritance pattern of their alleles. If the individuals in an afflicted family who develop the disease almost always inherit the same version of the marker, the mutant gene and the marker must lie very close together on the same chromosome. The marker and the disease gene are said to be linked.

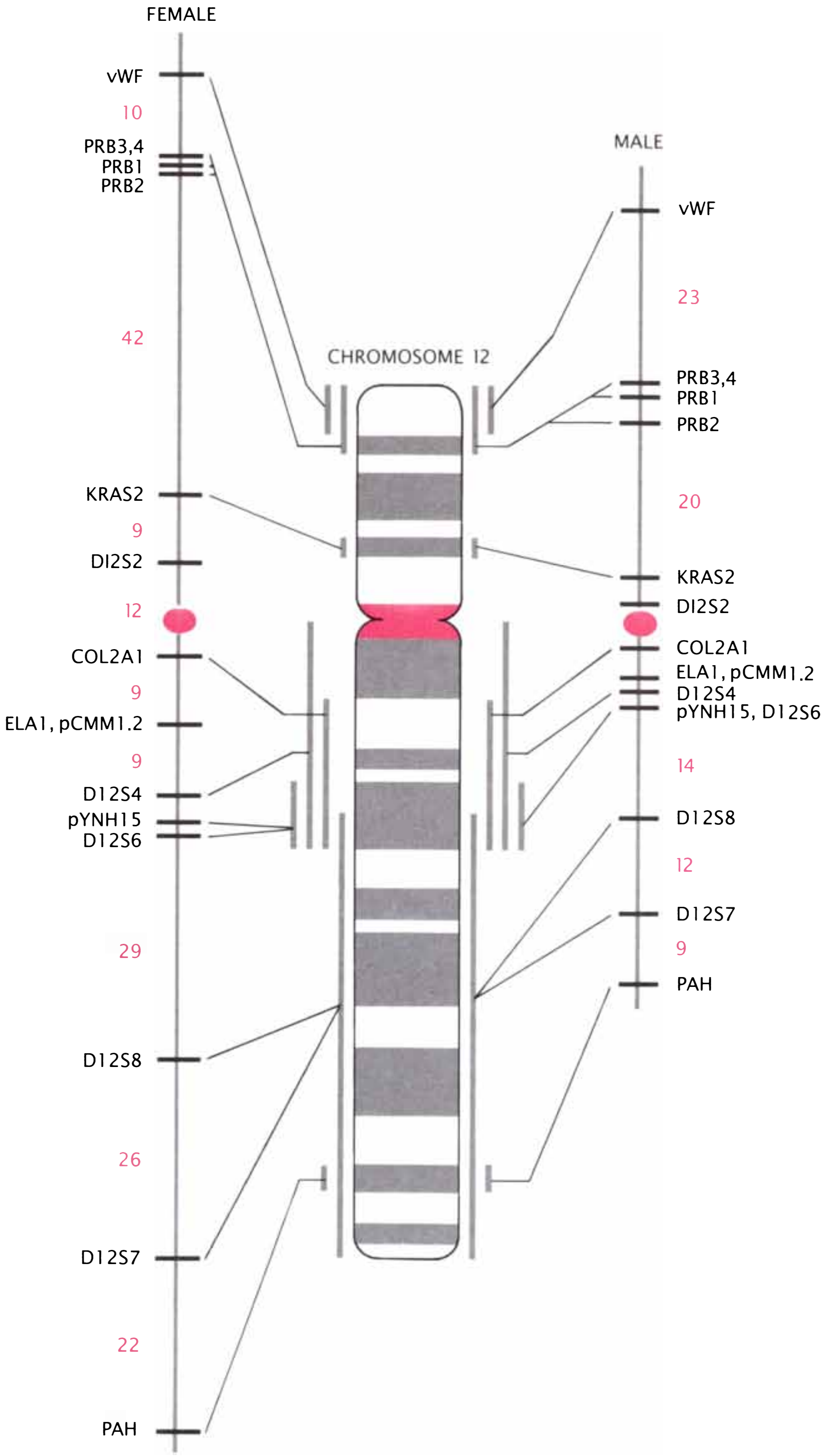
Other markers lying farther from the disease gene will recombine with the gene more frequently, so that the disease will be less likely to be inherited together with any given marker allele. In the extreme case, for a marker and a disease lying well apart on a chromosome, the recombination frequency reaches 50 percent. The marker and the gene are then unlinked: a given marker allele has only an even chance of being passed on with the disease. The same pattern of 50 percent coinheritance emerges when a marker and a mutant allele are borne on entirely different chromosomes.

Correlating the inheritance of a marker and a disease requires two things. The marker must be readily detectable, and it must be found in a number of distinguishable variants throughout the population. Linkage can be detected only if a person carrying mutant and normal alleles of a disease gene also carries two different versions of the marker; if the two marker alleles are indistinguishable, crossovers between the disease and the marker will be undetectable in the offspring. There will be no way to tell a linked marker from an unlinked one.

Until a few years ago only a limited set of markers met both criteria. The genes coding for certain enzymes, blood-group antigens (which determine blood type) and other proteins have multiple alleles, which manifest themselves by giving rise to protein polymorphisms: detectably different versions of the protein each gene codes for. Only 25 to 30 such marker systems of any value were known, however, covering only small sections of a few chromosomes. For want of markers most of the human genome remained inaccessible to the linkage approach.

With the advent of recombinant-DNA technology in the mid-1970's linkage mapping could be transformed into a practical and powerful tool for human genetics. The transformation can be dated to a genetics retreat sponsored by the University of Utah in April, 1978. There David Botstein of the Massachusetts Institute of Technology, Ronald W. Davis of Stanford University and Mark H. Skolnick of Utah proposed that the DNA sequence itself might yield numerous and readily detectable markers. Recognizing the potential power of the new approach, one of us (White) soon decided to test the hypothesis by committing his laboratory to the development of a set of DNA-based markers that would make it possible to detect linkage anywhere in the human genome. Botstein, White, Skolnick and Davis published the first paper detailing the approach in 1980. In the meantime

EXTENSIVE FAMILIES with living grandparents—modern counterparts to this turn-of-the-century family—are the ideal setting for studies of genetic linkage. In linkage studies the relative positions of sites in the chromosomes are inferred from the frequency with which genetic variations at those sites are passed on together from parents to children. By examining the inheritance of a genetic disease and arbitrary genetic markers in afflicted families one can assign a chromosomal location to the disease gene; by correlating inheritance of many markers in large, healthy families one can make maps of chromosomes.



MAP of chromosome 12 was made by tracing the inheritance of DNA markers: sites where the two copies of a chromosome often carry detectably different DNA sequences. The markers are arrayed in their statistically likeliest order and are separated by distances reflecting their recombination frequency, or the percent of the time marker versions carried on the same parental chromosome are separated by a recombination event during the formation of sperm or eggs. The recombination frequency between two markers rises with increasing physical separation, but the precise relation between recombination frequency and distance can vary depending on several factors, including sex. On chromosome 12, for example, the overall rate of recombination seen when the chromosome is passed on by a woman is higher than when it is passed on by a man, and so its genetic map is represented as being longer in women. An approximate chromosomal position has been determined for some of the DNA markers (*center*).

many other workers were beginning to find markers in human DNA and to speculate about their uses, and it was clear that this approach was an idea whose time had come.

The new linkage strategy gains its power from the very high level of normal polymorphism that can be found in the sequence of base pairs making up DNA. Between homologous chromosomes there is a difference in sequence, on the average, every 200 to 500 base pairs. Identifying these allelic variants would provide a practically limitless supply of markers scattered throughout the human chromosomes.

Molecular tools known as restriction enzymes provide a means of detection. Each restriction enzyme, made by a particular species of bacteria, binds to DNA wherever it finds a specific short sequence of base pairs and cleaves the molecule at a specific site within that sequence. A variation in DNA sequence that creates or eliminates a restriction site will alter the length of the resulting DNA fragment or fragments. The variation creates a restriction-fragment length polymorphism—an RFLP.

The RFLP defines a potential marker. A single restriction enzyme finds millions of cutting sites in the total human DNA, however. How can one or two variant fragments be detected among millions? The fragments are first sorted by electrophoresis: an electric field draws them through a gel, in which their mobility is inversely proportional to their length. A powerful and sensitive technique called Southern blotting after Edward M. Southern, who developed it at the University of Edinburgh, serves for picking out the fragments of interest.

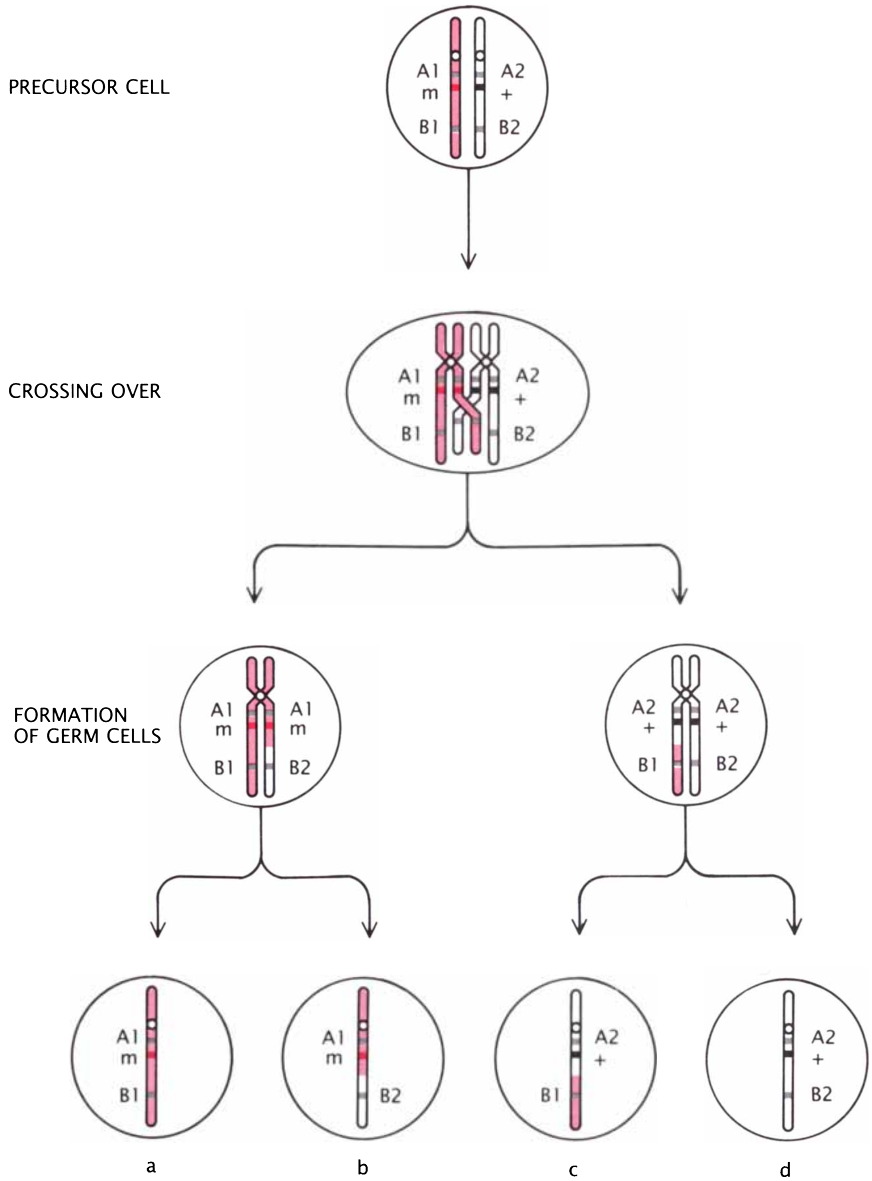
Southern blotting relies on the unique character of the DNA molecule. The bases along two strands of DNA can pair only according to set rules, and so the sequence on one strand constitutes a unique match for the sequence on the other. A length of single-strand DNA can therefore act as a probe, detecting and binding to the complementary sequence in a sample of ordinary DNA that has been "denatured": heated or exposed to high *p*H in order to separate its strands. In Southern blotting the DNA fragments on an electrophoresis gel are denatured and blotted onto a membrane, where they are exposed to probe DNA labeled with a radioactive isotope. The probe hybridizes, or binds, only to the fragment or fragments that bear the complementary sequence of bases. The radioactive label makes it possible to detect the position of the fragments, which reveals their size.

To detect an RFLP, then, one needs to find a probe that is complementary to DNA near the restriction-enzyme cutting site. A segment of DNA is chosen, often at random, from a collection (a "library") of cloned DNA fragments representing the full human genome. It is denatured, made radioactive and applied to Southern blots of DNA samples that have been digested with a restriction enzyme. If the radioactive bands appear at different places on blots of DNA from different individuals, the cloned DNA has detected the variable cutting pattern that results from a DNA polymorphism. The probe and the RFLP it detects constitute a unique genetic marker system. With it one gains a point of reference in the genome: the short stretch of polymorphic DNA, whose inheritance pattern can now be traced.

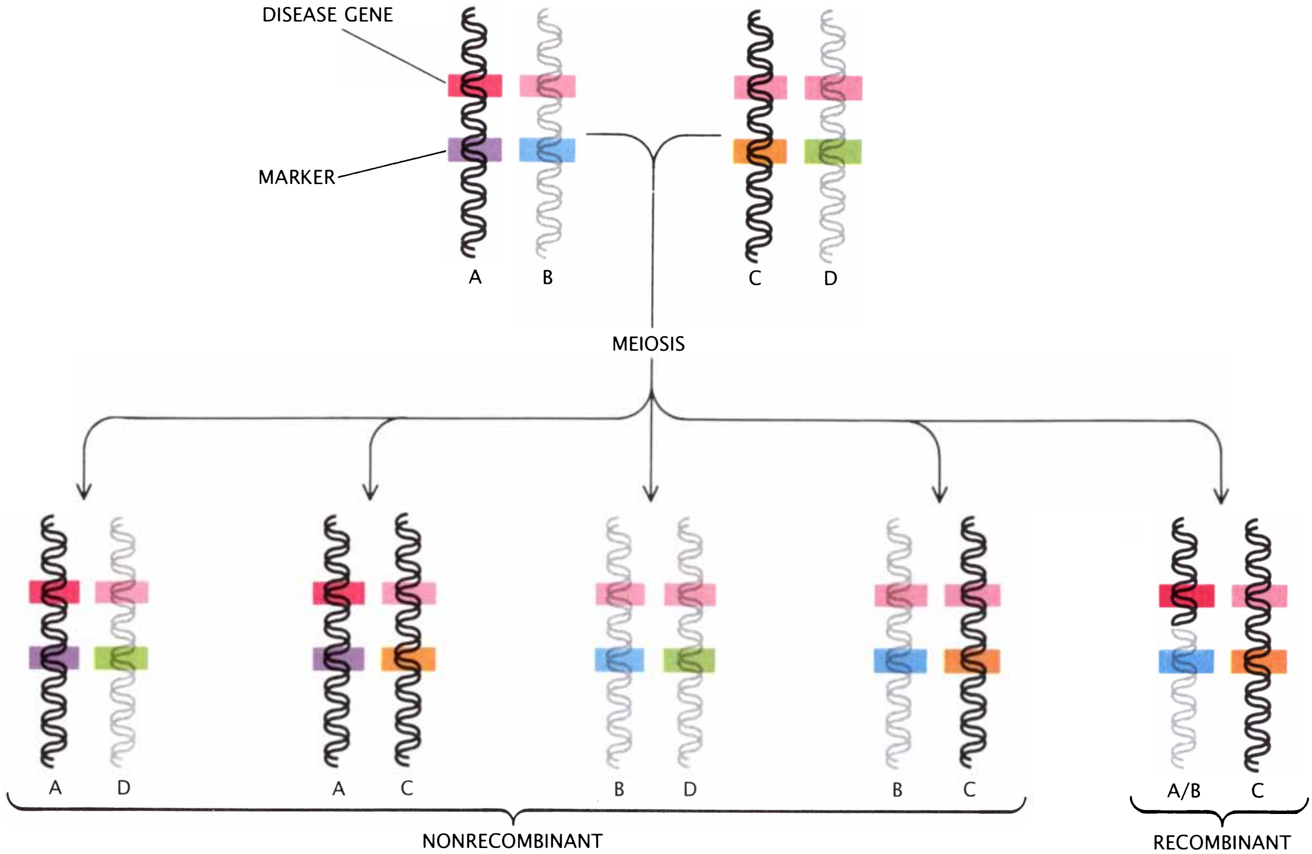
This DNA marker, defined by the RFLP, is found in one form or another in every individual, healthy or diseased. But if a genetic disease is passed down a pedigree together with a particular allele of the RFLP, the mutant gene can be assumed to lie in the same chromosomal region as the marker. In a second afflicted family the same marker will also show linkage, although the specific form of the marker that accompanies the disease may differ. Linkage to an arbitrary DNA marker reveals nothing about the physical position of the gene itself, of course, and for many purposes (such as diagnostic tests) physical location is immaterial. Nevertheless, the probe can also pick out the chromosome carrying the marker and the disease gene. If the probe is applied to a full set of human chromosomes, for example, it will hybridize to the chromosome bearing the marker site.

The value of any marker depends in large part on how many variants it displays throughout the population: the more versions of the marker there are, the more likely it is that an individual harboring a disease gene will carry two different alleles at the marker locus, making it feasible to detect recombination between the disease and the marker in offspring. Many RFLP's result from a change in a single base pair or the addition or deletion of a few base pairs at the restriction-enzyme cutting site. Such variation has a simple effect: the restriction site is either present or absent. The RFLP exists in only two forms, and so at least half of all individuals will probably be homozygous at the marker locus: they will carry the same variant on both homologous chromosomes. (Occasionally two restriction sites occur sufficiently close together to be detected by a single probe, yielding in effect a single marker with four alleles.)

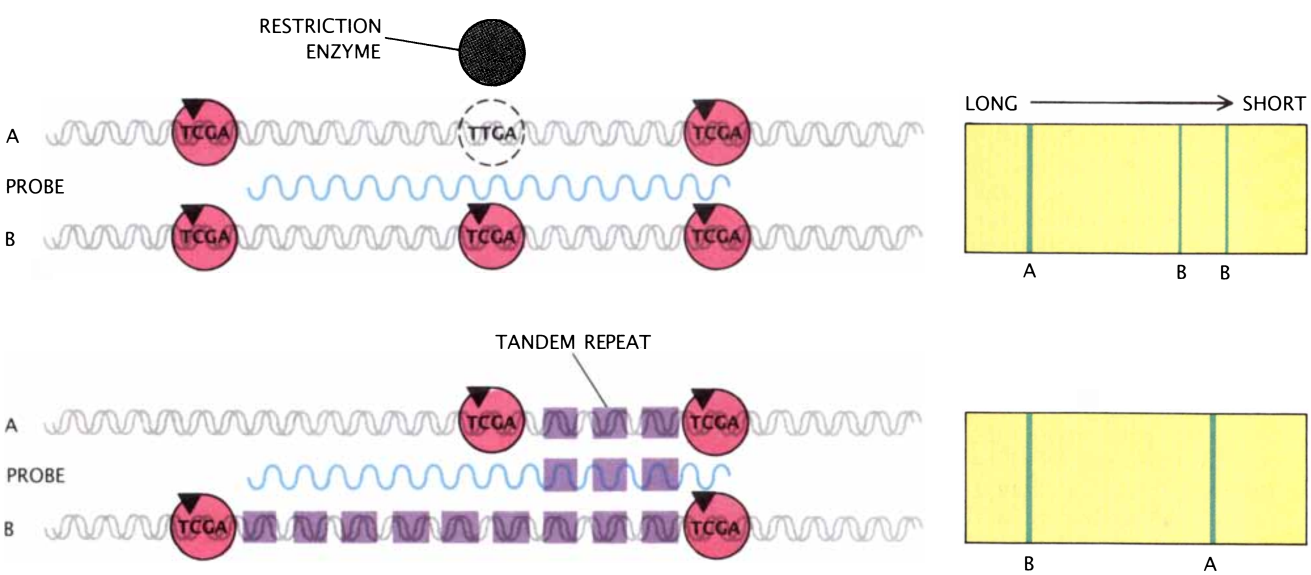
Another kind of DNA polymorphism creates many different versions of an RFLP. At many sites on the human DNA a single sequence that does not code for any protein is repeated many times. The origin and significance of these "tandem repeats" is a mystery, but for linkage mapping they offer a practical advantage in that the number of repeats at a given



RECOMBINATION makes it possible to detect genetic linkage. The diagram follows one idealized pair of homologous, or matching, chromosomes through meiosis, the process of cell division that produces germ cells (sperm or eggs). The chromosomes carry different alleles of two markers (*A, B*); one chromosome also bears a mutant, disease-causing allele of a gene (*m*) and the other chromosome bears the normal allele (+). In the precursor cell the disease is associated with allele *1* of both marker *A* and marker *B*. In the first phase of meiosis the chromosomes are replicated. The homologous chromosomes then "cross over," exchanging segments of equal length. Here crossing over takes place between loci *A* and *B*. The result is two germ cells (*a, d*) that carry the parental combinations of alleles and two (*b, c*) that contain recombinant chromosomes. In cell *b* the mutant gene is still found with allele *1* at locus *A* but is now joined by allele *2* at locus *B*. A low frequency of crossovers between the disease gene and marker *A* in many meioses would indicate that the disease and the genetic marker are closely linked.



LINKAGE between a disease gene and a marker is evident in the family history of the disease. Genetic features of a hypothetical couple and their children are shown. One parent suffers from a genetic disease caused by a single mutant allele (*red*); the other is healthy and hence carries only normal versions of the gene (*pink*). Children who inherit the disease usually also inherit a particular marker allele (*purple*) from the diseased parent, since the disease gene and the linked marker tend not to recombine.



DNA MARKERS—sites at which homologous chromosomes often differ in DNA sequence—are detected as RFLP's (restriction-fragment length polymorphisms). The DNA is digested with a restriction enzyme, which cuts wherever it finds a specific short sequence of nucleotides (in this case the base sequence *TCGA*). In one kind of marker (*top left*) a sequence difference causes a restriction site that is present on one homologous chromosome to be absent from the other. As a result the restriction fragments from each chromosome will differ in length. A DNA probe whose base sequence is complementary to that of DNA at the marker site reveals the fragments after they are sorted by electrophoresis (*top right*). Another kind of marker (*bottom left*) is characterized by a VNTR—a variation in the number of tandem repeats (short, repeated DNA sequences). The span between cutting sites differs between matching chromosomes, again resulting in distinctive fragments detected after electrophoresis (*bottom right*).

locus can vary from a few to hundreds of copies. Restriction fragments generated by cutting near these tandem repeats vary in length correspondingly [*see bottom illustration on opposite page*]; hence the RFLP comes in not just two forms but many. Given this variability in the population as a whole, the odds are good that a given individual will carry different versions of the RFLP on homologous chromosomes. A Southern blot will reveal two distinct fragments of different lengths, one from each homologous chromosome.

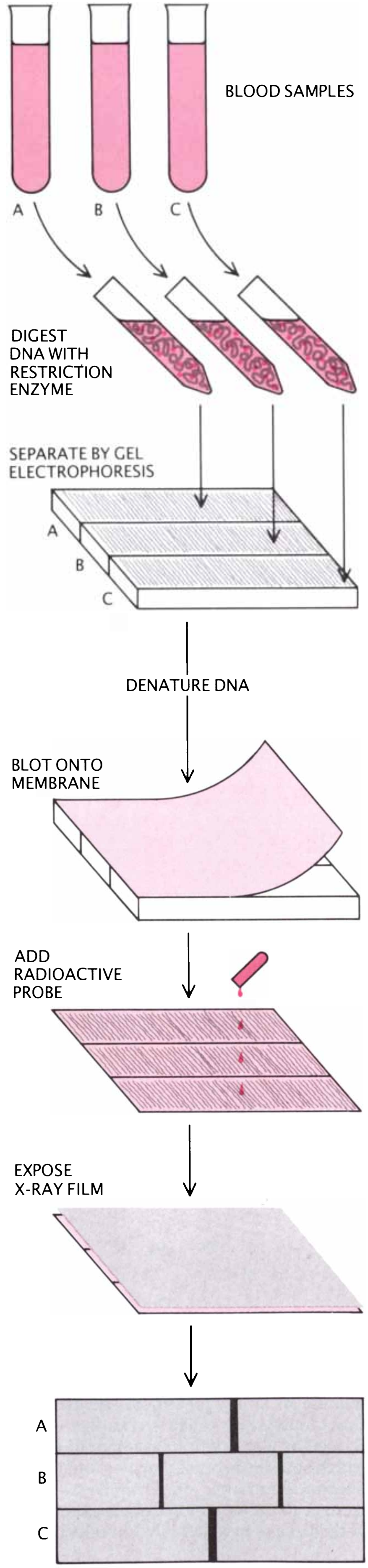
Probes for markers based on variations in the number of tandem repeats (VNTR's) can be developed more systematically than probes for ordinary markers. Alec J. Jeffreys of the University of Leicester recently recognized that the repeated sequences at many VNTR loci in different parts of the genome show similarities. The evolutionary explanation is again obscure, but the partial sequence homology means that under certain conditions a probe complementary to one VNTR locus can serve to pick out probes specific for other loci from a library of cloned DNA. Of the nearly 600 DNA markers developed so far in our laboratory, about 300 are VNTR's.

Such markers can serve as elements in an overall linkage map of the genome, or they can be developed for the more immediate purpose of tracking down a specific disease gene. Finding a marker whose inheritance is correlated with the appearance of a disease can be a staggering task on unmapped chromosomes. Since one often begins by knowing nothing about the chromosomal location of the disease gene or of any marker whose inheritance pattern is traced in an afflicted family, one can unwittingly search for linkage to tens of markers lying in a chromosomal region that is actually remote from the disease gene while leaving another, linked region unexamined. Nevertheless, the linkage strategy has already scored some remarkable successes.

If one knows which chromosome to search, the number of markers that must be tested can be reduced from several hundred, on the average, to as few as half a dozen. A genetic disease that almost always affects males but is inherited through the mother, for example, can be assumed to result from a recessive gene on the sex-determining *X* chromosome. (A mother carrying the disease has a second *X* chromosome bearing a normal copy of the gene, which masks the recessive disease gene; a son who inherits the mutation has only one *X* chromosome and therefore develops symptoms.) To find the gene, one need only test markers known to be carried on the *X* chromosome.

Genes for *X*-linked diseases were among the earliest to be traced through RFLP analysis; the first was the gene that causes Duchenne muscular dystrophy and probably also Becker muscular dystrophy (mapped by Kay Davies of the University of Oxford and Robert Williamson of St. Mary's Hospital in London). Yet an increasing number of diseases stemming from defects on the autosomes (the 22 pairs of nonsex chromosomes) have also yielded to the linkage strategy.

Huntington's disease became the first autosomal disease to be linked with a DNA marker when James F. Gusella of the Harvard Medical School and his colleagues studied afflicted families living in this country and near Lake Maracaibo in Venezuela. The group was fortunate in having to trace only eight markers through the families before finding one that was linked to the disease. Since then our laboratory and others have discovered markers for the genes causing disorders including cystic fibrosis, peripheral neurofibromatosis, or von Recklinghausen's disease (a disorder characterized by "café au lait" spots on the skin and a tendency to develop tumors and other disorders of the bone and nervous system), and familial polyposis coli (whose victims develop many colon polyps and run a very high risk of colon cancer).



RFLP ANALYSIS begins with a blood sample. DNA is extracted from the nuclei of white blood cells and digested with a restriction enzyme. The resulting DNA fragments are separated by gel electrophoresis, which sorts them in order of size. The RFLP is then detected by Southern blotting. First the DNA in the gel is heated to denature it, or separate its two strands, and is blotted onto a nylon membrane. A probe—a radioactively labeled segment of single-strand DNA that is complementary to the RFLP locus—is applied to the membrane. The probe hybridizes with the fragments from the locus; a sheet of X-ray film placed over the membrane detects the radioactively tagged fragments and thereby reveals which versions of the RFLP are present. In RFLP analysis of families, DNA samples from several individuals are often analyzed at the same time.

Tantalizingly, evidence of linkage has even been seen for forms of Alzheimer's disease and manic depression that run in families.

A "hit" can open the way to identifying the gene itself, which in turn provides a starting point for investigating the molecular mechanisms of the disease. By cloning the gene and determining its base-pair sequence one can deduce the composition of the protein it codes for and perhaps identify a specific defect. The protein can be synthesized and antibodies to the protein can be generated in experimental animals. Properly labeled, the antibodies can reveal the distribution of the protein in tissues affected by the disease. Such knowledge might hold the key to a treatment.

In many cases, however, the initial localization is too imprecise for a direct approach to the gene by current DNA technologies. The Huntington's disease gene, for example, recombines with its first identified marker at a frequency of about 5 percent, which implies that the marker lies as many as five million base pairs away from the gene. For identifying and cloning a gene the suspect stretch of DNA must be reduced to about a million base pairs, which means finding markers that recombine with the gene at a frequency of only about 1 percent. Ideally the markers will also flank the gene, bracketing the stretch of DNA to be tested.

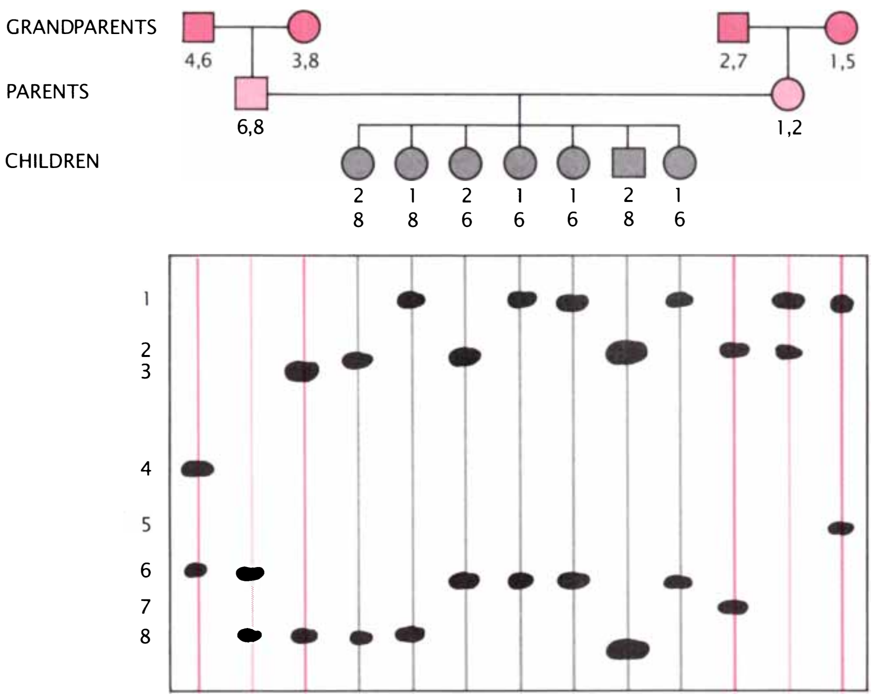
Tightly linked, flanking markers for cystic fibrosis, peripheral neurofibromatosis and familial polyposis are in hand, and a new, tightly linked marker has been identified for Huntington's disease. The search for the causative gene of each disease is in high gear. The approaches vary, but a common tactic is to comb a library of cloned chromosomal segments for one that is recognized by probes for both flanking markers. The segment—which presumably includes both markers and the gene they flank—can then be broken down further and each of the fragments cloned and tested for biological activity. Typically, a fragment can serve as a probe for messenger RNA (the sign that a gene is being expressed) in tissue affected by the disease. If it detects a messenger RNA that is unique to affected tissue, the probe itself may include all or part of the disease gene.

A different strategy has already culminated in the identification of the genetic defect in Duchenne muscular dystrophy. The region of the *X* chromosome that Davies and Williamson had linked with the disease shows missing segments in many patients; hence the disease may sometimes result from the absence of part or all of a normal gene. By identifying a region that is deleted in common among disease victims, Louis M. Kunkel of the Harvard Medical School and his colleagues were able to isolate and clone the gene.

Even before a disease gene is identified, linkage can sometimes point to possible causative mechanisms. The linked marker may fall near a gene of known function, which may then become a candidate for causing the disease. To take one instance, the marker for peripheral neurofibromatosis occurs on chromosome 17, which also carries the gene encoding the cellular receptor for nerve-growth factor (a substance that is vital for the survival and growth of nerve cells). That gene came under suspicion as a possible site of the mutation responsible for neurofibromatosis, but it was later found to lie some distance from the locus of the disease. Other genes on chromosome 17 may now become candidates for involvement in the disorder.

Reasonably tight linkage between a marker and a disease also makes it possible to devise tests for carriers and unborn victims—tests that are urgently needed given the frequency and insidious character of many genetic diseases. In populations of northern European descent, for example, one individual in 20 carries the cystic fibrosis gene. Because the gene is recessive, the carrier shows no symptoms, but if two carriers marry, their children stand a one-in-four chance of inheriting two defective genes and developing the disease. Huntington's disease is caused by a dominant gene (manifested even if the matching gene is normal), but its symptoms generally do not appear until middle age—after the unwitting victim has transmitted the disease to half of his or her children.

Before the presence of a disease gene can be established in an individual at risk, DNA from several other family members, both diseased and healthy, must be analyzed to determine which marker allele (or alleles,



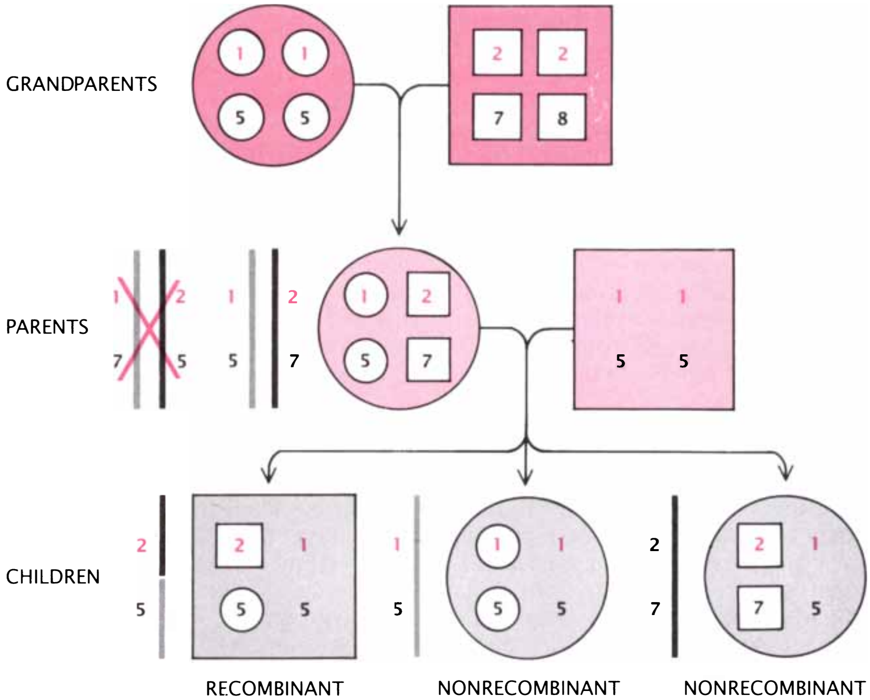
INHERITANCE OF AN RFLP can be traced by comparing restriction fragments from a number of family members. The RFLP marker that was analyzed in this three-generation family—a so-called VNTR locus—has many different alleles, each of them characterized by a restriction fragment of a specific size. Each individual in this pedigree (in which squares represent males and circles represent females) carries two different alleles of the marker, one from each homologous chromosome; children get one allele from each parent. If a particular allele of the RFLP is consistently associated with a genetic disease in an afflicted family, the marker and the defective gene may be linked.

in the case of a recessive disease that takes two copies of a gene to show itself) is inherited with the disease in this particular family. Finding a telltale allele in DNA from a prospective parent then indicates that he or she risks passing on the disease. Because DNA samples can be taken from a fetus soon after conception, the disease can also be diagnosed prenatally, enabling parents to make an informed decision about continuing the pregnancy. It is worth noting that in families at risk for some genetic diseases, fetal testing is actually increasing the number of births, simply because many couples would not conceive at all if they could not be sure the child was healthy before bringing it to term.

The construction of linkage maps, showing both arbitrary linkage markers and characterized genes arrayed along the chromosomes, has gone forward in parallel with the search for specific disease linkages. Linkage mapping represents a more deliberate and systematic approach to tracing mutant genes. From a complete linkage map workers trying to locate a disease gene will be able to choose and test a set of markers spaced at equal, large intervals along the chromosomes. Then, having discovered a linkage that restricts the gene's location to a specific chromosomal segment, they might test markers from a fine-scale map of the region in search of the tight linkage needed for further molecular studies.

The capacity to scan the genome for linkage not only will allow single-gene defects to be pinpointed more efficiently but also will hasten the search for the genetic bases of diseases caused by multiple aberrant genes. In addition, linkage maps will ultimately make it possible to check many points along the chromosomes simultaneously for a pattern of inheritance matching the family history of a disease, such as diabetes, coronary heart disease and certain cancers, to which susceptibility seems to be inherited. It might then be possible to close in on genes that confer predisposition to these illnesses.

Producing such a map extends the linkage strategy: now one is searching for linkage not between a DNA marker and a disease but between arbitrary DNA markers. Finding that alleles of different markers tend to be passed on together suggests the markers reside on the same chromosome, and the particular frequency with which the markers recombine reflects their "genetic distance."



DATA FROM THREE GENERATIONS can solve genetic mapping's "phase" problem, posed by two markers on the same chromosome. Unless one knows the phase of two markers (*color and black*) in a parent—how their alleles (*numbers*) are distributed between the homologous chromosomes—one cannot unambiguously detect recombination in the children. Analyzing DNA from grandparents (the mother's parents in this case) can reveal which two alleles each grandparent contributed. Since the mother must have received alleles *1* and *5* on the chromosome she inherited from her mother, alleles *2* and *7* can be assigned to the matching chromosome from her father. The other configuration of alleles is thus ruled out, and a recombination event that has taken place in the mother's chromosome can be identified unambiguously in the first child.

Although linkage mapping is simple in concept, it presents an enormous bookkeeping and analytical challenge. A large-scale linkage map of the genome, sufficient to locate any disease gene within a span of between 10 and 20 million base pairs, would include between 100 and 200 evenly spaced markers. To have markers at even intervals, however, one must assemble a much larger set of random markers on the map. DNA must be collected from hundreds of individuals in dozens of large families and tested for the RFLP's characterizing each marker locus.

The analysis of these vast data sets is complicated by the fact that perhaps two-thirds of the markers in any individual are uninformative. They carry two identical alleles, with the result that linkage between the marker and any other locus cannot be detected in offspring. For two markers that may be linked, moreover, there is often no way to determine their "phase": how their alleles are distributed between the two homologous chromosomes. Without knowledge of which alleles are on the same chromosome in a parent, one cannot unambiguously detect recombination between the markers in the child.

These limitations are minimized when the data are gathered from extensive pedigrees. We have been fortunate in being able to draw on excellent family resources for our own mapping effort. For one thing, more than 50 Utah families with eight children or more have generously volunteered to give blood samples, from which we take DNA for analysis and establish permanent cell lines. The presence of many children means that the parents' chromosomes can be followed through a large number of meioses, giving more accurate estimates of recombination frequencies than could be had from families with few children. In addition almost all the Utah families we sampled have four living grandparents, whose DNA can often indicate the phase of markers in the parents. If one knows, for example, that allele 1 of marker *A* and allele 3 of marker *B* both originated in a grandfather, then his child—one parent—must carry both alleles on the same chromosome if the markers are linked.

Even so, the inevitable limitations in the data mean that the map must be founded on probabilities. Statistical techniques make it possible to estimate the likeliest recombination frequency, and hence the genetic distance, between any two markers in the light of the observed inheritance pattern. An estimated recombination frequency of 50 percent suggests two markers are unlinked; a smaller frequency—say 10 percent—that has strong statistical support suggests linkage. Very early in our own mapping venture one of us (Lalouel, who was then at the University of Paris) realized that the task would demand specialized statistical methodology and computer programs. He and his colleague Mark Lathrop designed algorithms and programs capable of both maintaining the huge data base and performing joint analysis of the inheritance patterns of many markers.

Having identified a set of linked markers, one still needs to determine their order along the chromosome. In principle one could calculate the probability of each possible order's giving rise to the observed inheritance pattern and choose the likeliest arrangement. As few as 15 linked marker loci, however, can be arranged in 6.5 times 10^{11} different orders, an impossibly large number. In practice one can quickly eliminate entire families of improbable orders by considering loci in subsets of, say, three at a time.

For a flavor of the reasoning, suppose that in a large family specific alleles of linked markers *A, B* and *C* are usually passed on as a group: a child inherits all or none of them. In one child, however, the original alleles of *A* and *C* are inherited with another allele of *B;* in a second child the original allele of *B* is joined by other alleles of *A* and *C*. The sequence *A-B-C* is the least plausible sequence because it implies that double recombination—recombination both between *A* and *B* and between *B* and *C*—took place in both cases. (Under either alternative order, *A-C-B* or *B-A-C,* one recombination can explain each observation.)

We have designed computerized systems that employ such strategies to arrive at the likeliest order for an arbitrary number of linked markers. Once the most plausible order for a cluster of linked markers has been established, they can be assigned to a specific chromosome, for example by hybridizing one of the marker probes to a set of intact chromosomes. Linkage clusters are thereby assembled into a chromosome map.

The genetic distances on a chromosome's linkage map are related to physical distances (numbers of base pairs), but the relation is by no means direct. We have found, for instance, that the recombination frequency of a given pair of markers often differs significantly between the sexes. That is, the probability that two markers on a chromosome inherited from the mother will have recombined during her meiosis may be quite different from the probability of recombination between the markers on the same chromosome inherited from the father. On chromosome 13, for example, recombination frequencies are several times higher in females. On chromosome 11 the opposite is true in one interval, and in an adjacent interval the two sexes show similar recombination frequencies. The molecular basis for these intriguing variations is mysterious, but as a practical matter we have been preparing two maps of each chromosome, one map for each sex, showing identical marker orders but different genetic distances.

We have completed preliminary maps for most of the human chromosomes. Another group has recently published a similar collection of preliminary maps, based on a smaller number of reference families. Yet linkage mapping is an inherently collaborative enterprise: every group is looking for landmarks on the same terrain. Markers developed and studied in one laboratory often complement markers from another laboratory, in some cases bridging gaps between linked clusters.

A framework for cooperation has been set up by Jean Dausset at the Center for the Study of Human Polymorphism (CEPH) in Paris. The CEPH has undertaken to collect, store and distribute DNA from 40 families. The collection draws mostly on our Utah families but also includes DNA from families studied by other workers. Investigators from around the world (including our own group) get complete sets of DNA from the collection; in return workers report their markers and inheritance patterns to the CEPH, which makes the data available to all investigators and so lays the groundwork for a single genetic map.

The completion, probably within the next few years, of a high-resolution map will consummate the transformation of the human genome from uncharted territory to well-surveyed ground. Such a map can be expected to yield precise locations for most of the remaining well-characterized genetic diseases. A complete linkage map will also prove invaluable for guiding another large-scale investigation of the genome, which is still in the planning stage: an effort to determine the complete base-pair sequence of human DNA. Small islands of DNA along the chromosomes will most likely be sequenced first. The linkage markers within each island will serve to locate it in the larger landscape of the genome.

DISEASE	CHROMOSOME	DATE
HUNTINGTON'S DISEASE	4	1983
DUCHENNE MUSCULAR DYSTROPHY	*X* (GENE CLONED)	1983
POLYCYSTIC KIDNEY DISEASE	16	1985
CYSTIC FIBROSIS	7	1985
CHRONIC GRANULOMATOUS DISEASE	*X* (GENE CLONED)	1985
PERIPHERAL NEUROFIBROMATOSIS	17	1987
CENTRAL NEUROFIBROMATOSIS	22	1987
FAMILIAL POLYPOSIS COLI	5	1987
MULTIPLE ENDOCRINE NEOPLASIA IIa	10	1987

TABLE OF DISORDERS gives a small sample of the genetic diseases for which the chromosomal location of the defective gene has been determined with the help of linkage studies. The table also indicates which chromosome carries the gene and the linked marker and gives the year linkage was first reported. Reasonably tight linkage can make the marker useful for diagnosing the disease in members of an afflicted family; very tight linkage can open the way to identification and cloning of the disease gene.

It's Lufthansa. And we fly to more countries around the world than any other airline.

From 13 convenient gateways in the U.S., we offer more than 100 flights a week to the heart of Europe. Plus fast, easy Lufthansa connections to more than 130 cities beyond.

Maybe that's why travelers to Europe, to the Middle East, to the Far East, to Africa, and to South America feel at home on Lufthansa.

People expect the world of us.

Lufthansa
German Airlines

Lufthansa is a participant in the Frequent Flyer mileage programs of United Airlines, Delta, and USAir. See your Travel Agent for details.

A Flaw in a Universal Mirror

Without a slight asymmetry in a "mirror" called CP *invariance the universe as we know it would not exist; instead it would be devoid of matter. What force in nature causes the flaw?*

by Robert K. Adair

Why is there matter in the universe? If the approximate symmetry between matter and antimatter that has been observed were perfect, the universe would be elegantly simple but virtually empty of matter and of creatures made up of that matter who could contemplate that elegance. The existence of the universe as we know it comes from a flaw in a symmetry exhibited by a universal "mirror" called the *CP* mirror: a symmetry that says the outcomes of some events in nature should remain the same on changing matter to antimatter (*C*) and viewing the result in a mirror (*P*).

It seems that, at a time somewhere before the first millionth of a second after the universe was born in the fiery ball known as the big bang, matter and antimatter probably existed in equal amounts. There were almost exactly equal numbers of particles and antiparticles, all in thermodynamic equilibrium under conditions of enormous pressure and temperature. (For each particle there exists an antiparticle that has the same mass but opposite electrical properties; the antielectron, or positron, for instance, is the antiparticle of the electron.) Then, as the universe expanded and cooled, most of the particles found their corresponding antiparticles and the pairs annihilated each other. If the *CP* symmetry were exact, only the very few particles and antiparticles that had by chance not found annihilation partners would be left. But the symmetry was slightly flawed, and an excess of about one in a billion protons and one in a billion electrons survived to form, in the fullness of time, galaxies, stars, planets and ourselves.

In spite of the fact that this *CP*-invariance violation must lie at the center of the riddle of our existence, and in spite of extensive theoretical and experimental studies devoted to *CP*, little more is known about *CP* asymmetry today than was known a quarter of a century ago, when the flaw was first discovered. New theoretical ideas developed in the past few years, however, have driven a new generation of experiments—one of which was undertaken by my colleagues and me at the Brookhaven National Laboratory—that promise to throw new light on the flaw. These experiments are conceived to examine the character of the force responsible for the violation of *CP* symmetry and to determine if the very small asymmetry follows from a weak (milliweak) force acting twice or a very weak (superweak) force acting once in particle interactions.

A full appreciation of the slight but essential asymmetry in the *CP* mirror requires some insight into the important role of symmetry in the design of the universe. Most physicists today believe in a grand equation that defines the properties of the elementary forces and particles and think the equation must express fundamental symmetries of the universe. Although the detailed structure of the equation is not yet known, many of the important symmetries of the equation are well understood and can be readily grasped.

As Amalie Emma Noether of the University of Göttingen demonstrated in the early years of this century, symmetries of the fundamental equation are generally connected to conservation laws. Hence the observations that momentum, energy and angular momentum are conserved demonstrate that the grand equation does not differentiate between places, times and directions. It is natural to ask about further symmetries. In particular, does the equation differentiate between left- and right-handedness (mirror images) and thereby define a screw sense? Does the equation differentiate between particles and antiparticles?

Answers to these questions can be found through laboratory experiment. I shall illustrate the character of the investigation of fundamental symmetries with the help of Lewis Carroll's intrepid heroine Alice. And, properly, we begin with a study of symmetry under reflection in a looking glass. The looking glass is more formally called a *P* mirror. (The *P* stands for parity.) Can Alice, perhaps awakening from a Carroll dream, tell whether she is in the looking-glass world or in her home world? Alice can be certain of her whereabouts only if she finds some fundamental process or structure that defines a screw direction. A right-hand screw at home will be a left-hand screw through the looking glass.

Until 1956 physicists would have replied that Alice must remain lost between the real world and the looking-glass world. The reason is that at the time it was generally believed that none of the fundamental interactions—gravity, electromagnetism, the strong interactions (which are responsible for nuclear forces) and the weak interactions (which are responsible for certain kinds of nuclear decay)—define a screw sense. In other words, it was held that the *P* mirror was symmetric, so that one could not distinguish between the result of a reaction among elementary particles and the result of the mirror image of that reaction.

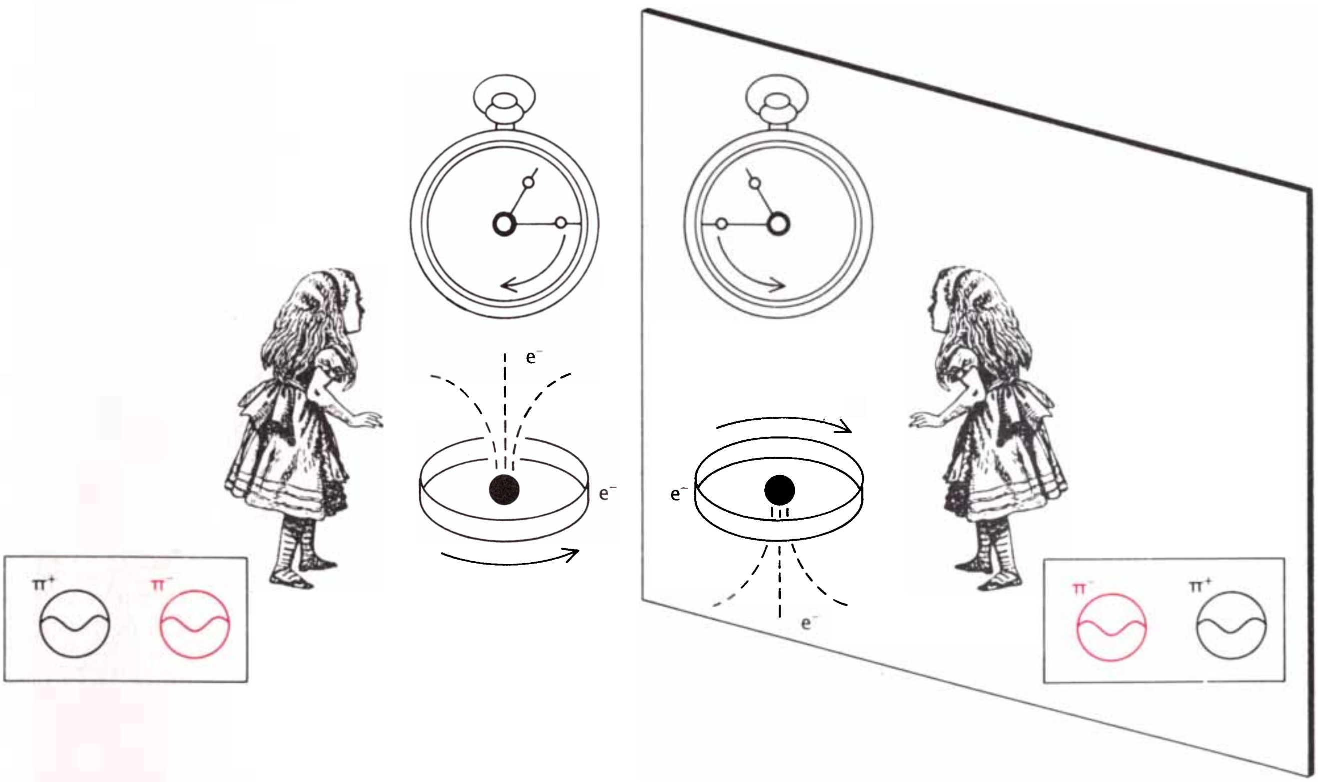
Then Richard H. Dalitz, working at Cornell University, demonstrated that properties of the decay of a particle called the *K* meson through the weak interaction appeared to be inconsistent with the notion that nature does in fact differentiate between left- and right-handedness. In the course of efforts to understand this, Tsung Dao Lee and Chen-Ning Yang proved in 1956 that there is nothing contradictory about supposing the weak interactions define a screw direction. In essence, Lee and Yang's theory, for which they were awarded a Nobel prize in 1957, shows that the *P* mirror could be asymmetric with respect to the weak interactions. In December, 1956, Chien Shiung Wu of Columbia University, Ernest Ambler of the National Bureau of Standards and other collaborators did an experiment suggested by Lee and Yang. The experiment showed conclusively that the weak interactions do define a screw direction: nature discriminates between left- and right-handedness.

Consequently Alice can tell whether she is in the looking-glass world or in her home world by doing an experiment along the lines of Wu and Ambler's. She places a cylinder of cobalt 60, an isotope of cobalt that decays by emitting electrons through the weak interactions, in a magnetic field generated by electrons flowing counterclockwise around a circular wire loop [*see left half of illustration below*]. Following elementary laws of physics, the magnetic field generated by the electrons moving in the loop points downward. Now, the nuclei of all atoms have a certain spin, or angular momentum, that causes them to act like tiny bar magnets spinning about their axes. In the case of cobalt nuclei the magnetic field aligns the nuclear spins downward, in the direction of the field.

Alice, in her home laboratory, finds that the electrons emitted by cobalt through the weak decays travel upward, in the direction opposite to the nuclear spins. Alice-through-the-looking-glass [*see right half of illustration below*] performs the same experiment, but because the electrons in the wire loop in the looking-glass world move clockwise, they generate a magnetic field that points upward and aligns the spins of the cobalt nuclei upward. The electrons emitted by the cobalt nuclei in the looking glass therefore travel downward. That means the real world and the looking-glass world are not absolute reflections of each other. The symmetry of the *P* mirror is broken.

If Charles Dodgson had been born a century later, he might have created another mirror for Alice and different adventures. Perhaps she might have passed through a *C* mirror, which would transform particles into antiparticles. (The *C* stands for charge conjugation.)

In the late 1920's the British physicist P. A. M. Dirac, who first postulated the existence of antimatter, showed that matter and antimatter must behave the same way under electromagnetic forces. Later it was seen that matter and antimatter are acted on identically by gravity and also by the strong nuclear forces. If the symmetry holds for all interactions, Alice can make no observations that will tell her whether she has passed through the *C* mirror.



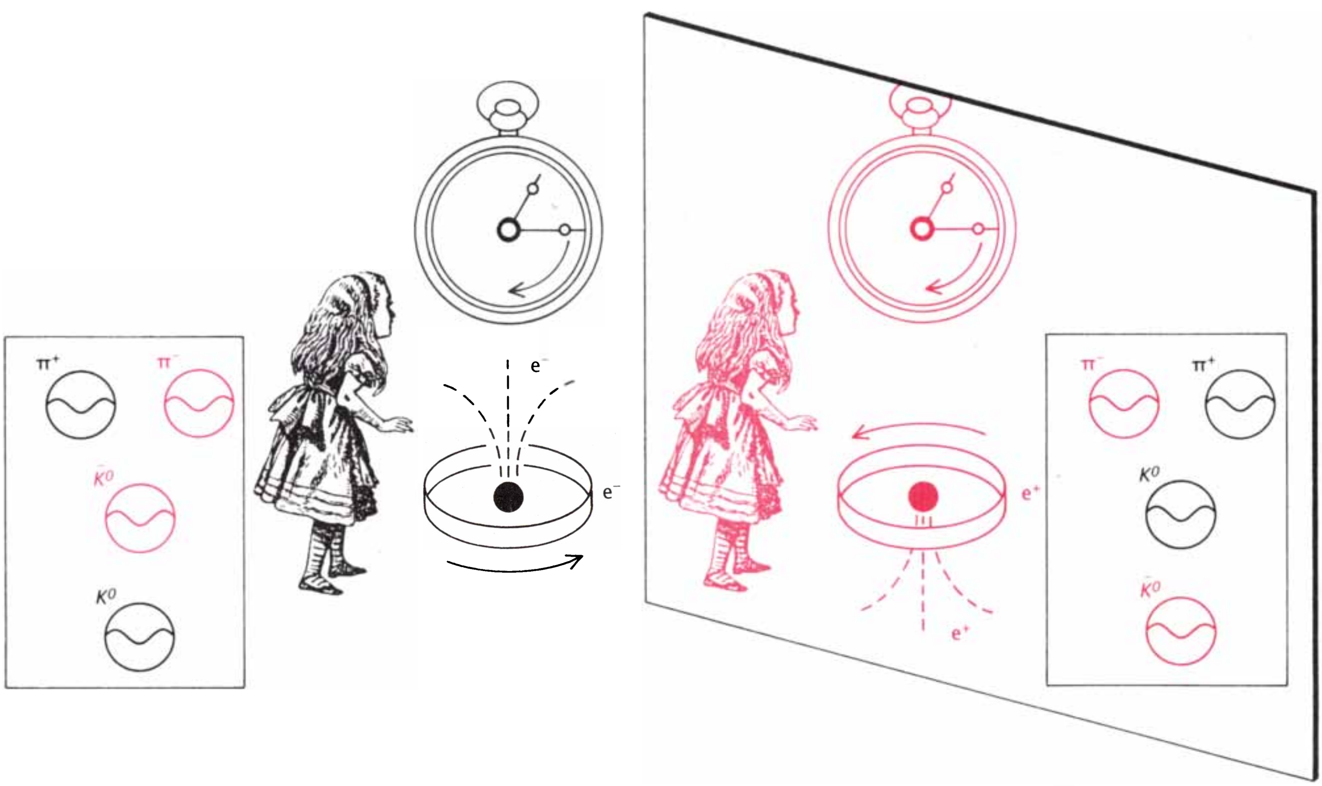
ALICE-IN-WONDERLAND SCENARIO provides a way of understanding symmetry laws in nature. Shown here is a test for reflection symmetry in a looking glass, which is more formally called a *P* mirror. (The *P* stands for parity.) Can Alice tell whether she is in the real world (*left*) or the looking-glass world (*right*)? To find out, she places a cylinder of cobalt 60, an isotope of cobalt that decays by emitting electrons, in a magnetic field generated by electrons flowing counterclockwise around a circular wire loop. The field points downward and aligns the spins, or angular momenta, of the cobalt nuclei downward; the emitted electrons travel upward. In the looking-glass world the electrons in the circular loop travel clockwise, and so the magnetic field points upward. The spins of the looking-glass nuclei align upward and the emitted electrons travel downward. Since the emitted electrons travel in opposite directions in the two worlds, the reflection is not perfect: the symmetry of the *P* mirror is broken. The *P* mirror does preserve the identities of particles and antiparticles (*panels*). An antiparticle (here the negatively charged π^- meson) has the same mass as its corresponding particle (in this case the positively charged π^+ meson) but opposite electrical properties.

The experimental and theoretical work connected with the discovery of the asymmetry of the *P* mirror, however, also showed that the weak interactions distinguish between matter and antimatter: the symmetry of the *C* mirror is also broken. In particular, aligned anticobalt nuclei emit positrons (the commoner name of antielectrons) preferentially in the direction of alignment.

Alice can then find out whether she has passed through the *C* mirror by determining either the direction of emission of the electrons emitted by her aligned cobalt nuclei or, alternatively, the direction of emission of the positrons emitted by her aligned anticobalt nuclei. Once again, at home electrons traveling counterclockwise in a wire loop generate a magnetic field that points downward; the cobalt nuclei align downward and the electrons they emit move upward [*see left half of illustration below*]. In the antiworld seen through the *C* mirror, positrons traveling in the wire loop also move counterclockwise (since the *C* mirror reverses charge), but because they have a charge opposite to that of electrons, the magnetic field points upward. The anticobalt nuclei have magnetic properties opposite to those of ordinary cobalt nuclei, and so their spins align downward. Finally, the positrons emitted by the anticobalt nuclei travel in the direction of the nuclear spins, which in this case is downward [*see right half of illustration below*]. The fact that the emitted positrons travel in the direction opposite to that of the emitted electrons tells Alice she must have passed through the *C* mirror to the antiworld.

I alluded above to the importance physicists place on finding symmetries in nature. One can therefore well imagine that many physicists were upset by the asymmetry of the *P* mirror. I remember feeling that I no longer could hold anything I knew as being certain. The solid earth on which I stood had turned to quicksand. But with the observation that the asymmetry in the *P* mirror always seemed to be accompanied by a compensating asymmetry in the *C* mirror, so that everything was symmetric under *CP*, the universe seemed to me, and to other physicists, to regain much of the order we thought had been lost. In other words, if Alice passed through the *CP* mirror, she would observe the same results as if she were at home [*see illustration on opposite page*]. For about seven years the view that the elementary forces were absolutely symmetric under *CP* survived.

Then in 1964 James Cronin and Val L. Fitch of Princeton University, working at Brookhaven, showed that small violations of *CP* symmetry do occur. I shall discuss their experiment below and consider here a related result that is easier to conceptualize. If Alice accelerates a beam of positively charged K^+ mesons to high energies and directs the particles into a block of copper, a beam of neutral *K* mesons will emerge. She finds that far from the target the neutral *K* mesons decay into positrons slightly more often than they do into electrons. Through the *CP* mirror Alice



***C* MIRROR changes particles to antiparticles and antiparticles to particles but preserves the relative orientations of the objects it "reflects." (The *C* stands for charge conjugation.) At the left, once again electrons traveling counterclockwise in a wire loop produce a magnetic field pointing downward; the cobalt nuclei in the cylinder emit electrons that travel upward. In the antiworld seen through the *C* mirror (*right*) positrons, or antielectrons, also travel counterclockwise through the loop (since the *C* mirror changes charge only), but because they have a charge opposite to that of electrons, the magnetic field points upward. Anticobalt nuclei have magnetic properties opposite to those of ordinary cobalt nuclei, and so their spins align downward. The anticobalt nuclei emit positrons in the direction of their nuclear spins, so that the positrons travel downward. Since the emitted positrons travel in the direction opposite to that of the emitted electrons, the symmetry of the *C* mirror is broken. The panels show that the *C* mirror changes the charge of π^+ and π^- mesons and changes the K^0 meson into its antiparticle, the $\bar{K}^0$ meson.**

accelerates a beam of negatively charged K^- mesons (the antiparticles of K^+ mesons) into a block of anticopper. Again neutral K mesons emerge, which also decay into positrons slightly more often than they do into electrons. Because in both cases the neutral K mesons decay into positrons more often than they do into electrons, CP symmetry is broken and Alice can determine whether she is home or has passed through both the P mirror and the C mirror.

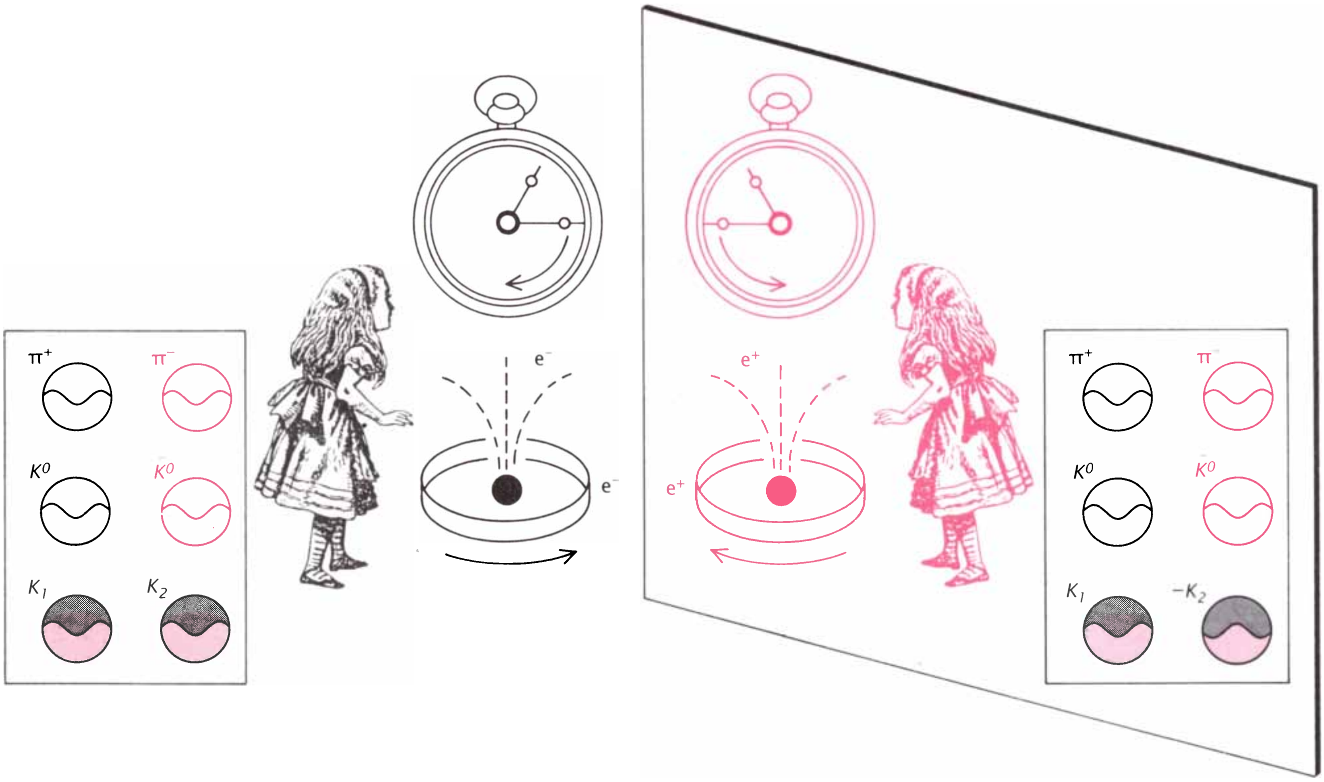
To make the determination Alice relies on the fact that a magnetic field will deflect a charged particle moving through the field. She proceeds in her quest by measuring the deflection of the positively charged positrons and negatively charged electrons that the neutral K mesons decay into. At home, once again, the magnetic field is produced by electrons flowing counterclockwise through a circular loop, so that the field points downward. Through the CP mirror the magnetic field is produced by positrons flowing clockwise, and so there the field also points downward. Alice finds that the chief decay products (positrons) from the neutral K mesons are deflected toward her, whereas Alice-in-CP-world finds that the decay products are deflected away from her [*see illustration on next page*]. By observing the deflection she can tell where she is. The experiment shows that not everything is invariant under the combination of reflection in a P mirror and the change of particle to antiparticle induced by a C mirror.

The experiment demonstrates a deep and fundamental difference between matter and antimatter and sweeps away the necessity imposed by the purported CP symmetry that the universe must contain no excess of matter. We do not yet know the exact character of the mechanisms that lead to the extra matter from which our world and we are made, but we know from Alice's experiment that our existence is not in contradiction with the symmetries we see.

A more complete understanding of CP asymmetry requires a more thorough understanding of the particles that are subject to the asymmetry. Why does the neutral K meson behave the way it does?

Up to this point I have, for the sake of simplicity, referred to the neutral K meson as if it were a single kind of particle. In reality the world is more complicated. First, there is the neutral K meson designated K^0. Then there is that particle's antiparticle, the $\bar{K}^0$ meson. The bar over the K simply signifies antiparticle.

A neutral K meson need not be either a K^0 or a $\bar{K}^0$ meson; it can be a mixture of the two. The concept of such mixing seems strange when one thinks of the mesons as particles like stones or balls but is less curious if one thinks of the mesons in terms of their description in quantum mechanics as (de Broglie) waves with properties much like water waves or light waves. Even as one can construct new waves from mixtures of the amplitudes of two different waves, one can construct new neutral K mesons from the amplitudes of the K^0 and $\bar{K}^0$ mesons.



COMBINED *C* MIRROR AND *P* MIRROR—a *CP* mirror—was held by physicists to be unflawed until 1964. It was thought the asymmetry of the *P* mirror would always be accompanied by the asymmetry of the *C* mirror, so that a *CP* mirror would be symmetric. In other words, if Alice were to pass through the *CP* mirror, the results of any experiment she did would be the same. In the special case shown here the results are in fact the same: both the electrons (*left*) and the positrons (*right*) are emitted upward. The bottom panels indicate the effect of the *CP* mirror on various particles, including the neutrally charged K_1 and K_2 mesons, each of which is made up of equal amounts of the K^0 and $\bar{K}^0$ mesons. (The "mixing" of particles is one of the strange phenomena predicted by quantum mechanics.) Every particle has associated with it a wave function, with properties very much like a water wave or a light wave, that describes the properties of that particle. The wave function associated with the K_1 meson is symmetric with respect to the *CP* mirror and the wave function of the K_2 meson is antisymmetric (hence the minus sign on the right).

Although neutral K mesons can be constructed with any degree of mixing, for our purpose we need consider only the K_1 meson, which is made by adding equal K^0 and $\bar{K}^0$ amplitudes, and the K_2 meson, which is made by subtracting a $\bar{K}^0$ amplitude from an equal K^0 amplitude. Roughly speaking, the K_1 and K_2 mesons are half particle and half antiparticle.

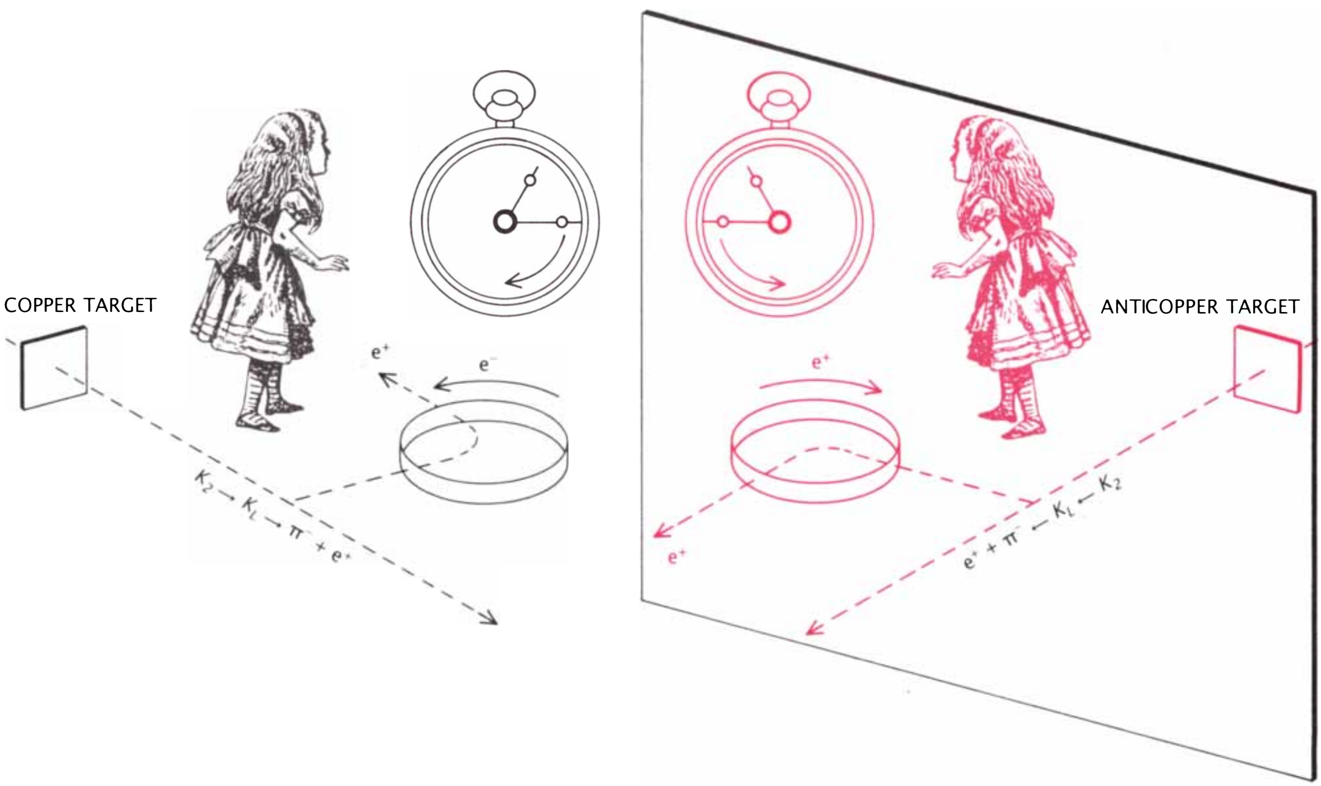
There is no a priori reason to think of the K^0 and $\bar{K}^0$ mesons as "pure" particles and the K_1 and K_2 mesons as mixtures. We can just as well think of the K_1 and K_2 mesons as basic and the K^0 and $\bar{K}^0$ mesons as mixtures. The K^0 meson is the sum of the K_1 and K_2 ($K^0 = K_1 + K_2$), and the $\bar{K}^0$ meson is the difference ($\bar{K}^0 = K_1 - K_2$).

When Alice bombarded her copper target with the positively charged K^+ mesons, the neutrally charged K^0 mesons emerged. When Alice-in-*CP*-world bombarded her anticopper target with the negatively charged K^- mesons, the neutrally charged $\bar{K}^0$ mesons emerged. I have shown that each of these initial beams of K^0 and $\bar{K}^0$ mesons can be thought of as a mixture of equal amounts of K_1 and K_2 mesons. The K_1 component of both beams decays rather quickly into pairs of particles called π mesons, leaving just the K_2 component.

Since the K^0 part of the K_2 meson decays into positrons and the equally large $\bar{K}^0$ part decays into electrons, there is no excess of positrons over electrons from K_2 decay, and an observation of that decay cannot tell Alice where she is. The *CP*-invariance-violation flaw in the mirror, however, acts to mix spontaneously a very small K_1 part into the K_2 to form a combination called the K_L, and the K_L, with a slightly larger K^0 than $\bar{K}^0$ part, decays more often into positrons than into electrons. (The L stands for long-lived.) It was this effect of the *CP*-invariance violation that Alice used to find her place.

The *CP*-invariance-violating addition of a K_1 part to the K_2 to form the K_L is evident in a second way. As Cronin and Fitch observed in 1964, the K_L decays into two π mesons. This decay follows from the K_1 part of the K_L; the K_2 virtually never decays into two π mesons.

The difference in the way the K_1 and K_2 mesons decay is rooted in the *CP* properties of the particles. On reflection in a perfect *CP* mirror the wave amplitude of the K_1 is unchanged, as is the amplitude of the wave describing two π mesons, but the amplitude of the K_2 is reversed, or changes sign. Consequently the K_1 decays quickly into pairs of π^+ and π^- mesons and pairs of π^0 mesons, whereas the K_2 does not. (The K_2 does



FLAW IN *CP* MIRROR is demonstrated by Alice. She accelerates a beam of positively charged K^+ mesons into a block of copper (*left*), causing a beam of neutrally charged K^0 mesons to emerge. The K_1 component of the K^0 mesons decays quickly into pairs of π mesons, leaving the K_2 component, which after traveling for some time changes to a particle called the K_L meson by adding a tiny amount of K_1 component. This K_L combination decays into positrons more often than it decays into electrons. Alice-through-the-*CP*-mirror in the meantime accelerates a beam of negatively charged K^- mesons into a block of anticopper (*right*), causing a beam of neutrally charged $\bar{K}^0$ mesons to emerge. The K_1 component of the $\bar{K}^0$ mesons decays quickly into pairs of π mesons, again leaving the K_2 component. After traveling some distance the K_2 meson again changes to the K_L meson, which in turn decays into positrons more often than it decays into electrons. Because in both cases the particles that were initially neutral K_2 mesons decay into positrons more often than they do into electrons, *CP* symmetry is broken. Alice can determine whether she has passed through the *CP* mirror by making use of the fact that a magnetic field will deflect a charged particle moving at a right angle through the field. In the real world, once again a magnetic field is produced by electrons flowing counterclockwise through a circular loop, so that the field points downward. Through the *CP* mirror positrons flowing clockwise generate a field that also points downward. Alice finds that the positrons from the K_2 mesons are deflected toward her, whereas Alice-through-the-*CP*-mirror finds that the positrons are deflected away from her. The symmetry of the *CP* mirror is thus broken.

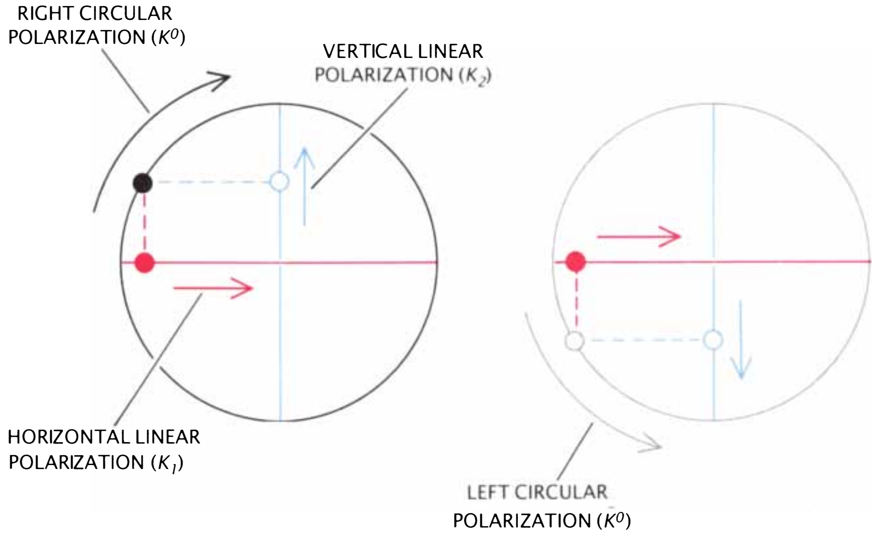
decay in other ways but not quickly.)

Like Alice, Cronin and Fitch started with a beam of K^0 mesons [*see illustration on next page*]. Again the K_1 component of the beam decayed quickly into pairs of π mesons, leaving only the K_2 component. If the *CP* mirror were perfect, the *K* meson would remain a K_2 meson and would not decay into pairs of π mesons. The *CP* asymmetry, however, acts so that the K_2 spontaneously adds a very small amount of K_1 amplitude and, through that amplitude, about one out of every 500 of the K_L mesons decays into a pair of π mesons.

Is the *CP* mirror broken for systems that do not involve the *K* meson? Although physicists are quite certain that the flaw in the *CP* mirror is universal, the magnitude of the effects seems to have been too small to have been seen in other systems that have been investigated. Perhaps, then, a new, very weak force is responsible for the effect. What would be the character of this force?

I have noted that the *CP* violation observed in *K* mesons can be described in terms of the addition of a small K_1 part to a K_2 meson. Phrased somewhat differently, a small part of the $\bar{K}^0$ meson, which is part of the K_2, turns into a K^0 meson.

Mesons are known to be compounds of a particle called a quark and an antiquark. *K* mesons, for instance, are made up of two kinds of quarks called down and strange quarks and their corresponding antiquarks, antidown and antistrange; the K^0 meson consists of a down quark and an antistrange quark, and the $\bar{K}^0$ meson consists of a strange quark and an antidown quark. For the $\bar{K}^0$ meson to change into a K^0 meson two events must take place: an antidown quark must change into an antistrange quark and a strange quark must change into a down quark. (Quarks do not change into antiquarks.) Such quark changes are known to be induced by weak forces, but they conserve *CP*, so that a change in one direction ($\bar{K}^0$ to K^0) is no more probable than the reverse change (K^0 to $\bar{K}^0$). There is no net change: the *CP* mirror is not broken.

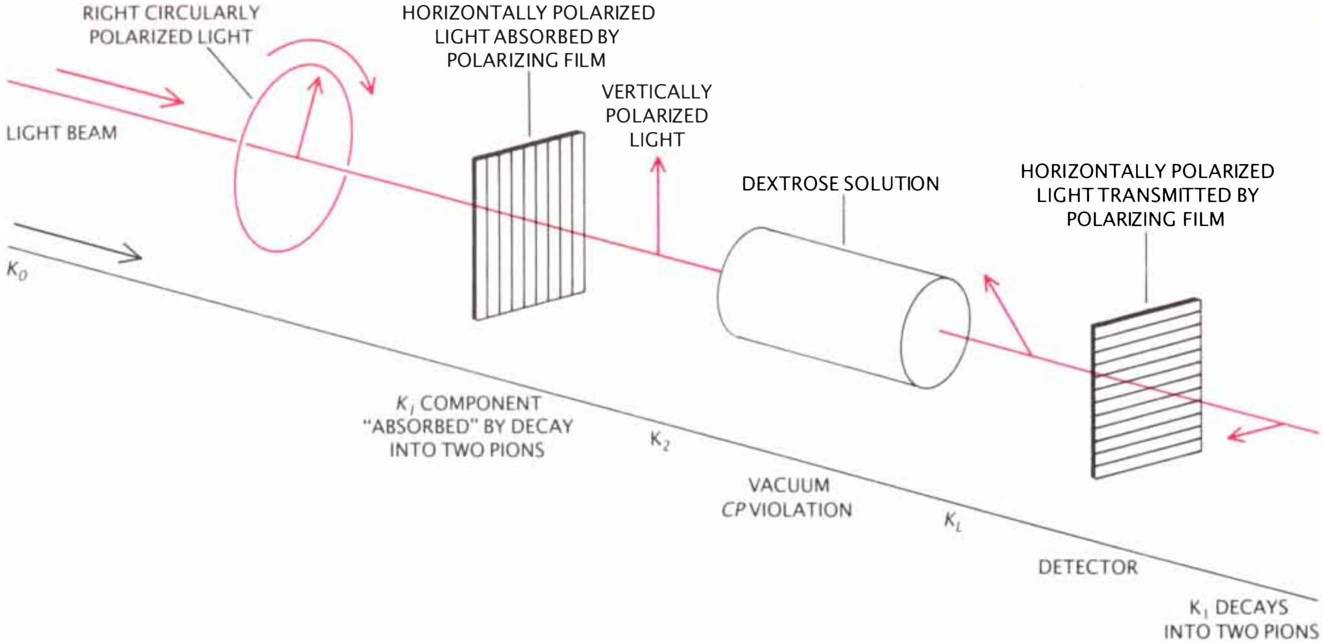


ANALOGY between properties of light and neutral *K* mesons can help to explain the "mixing" of those mesons. Light can be polarized in different ways. Suppose a beam of light points out of the page. If the oscillating electric fields of that beam are directed left and right in the plane of the page, the light is said to be horizontally polarized; the K_1 meson can be considered to correspond to such a polarization. If the electric fields are directed up and down in the plane of the page, the beam is vertically polarized; the K_2 meson corresponds to that polarization. If the electric field direction moves circularly, much as the direction of the wing on a wing nut changes as the nut advances on a threaded bolt, the light is said to be circularly polarized. If the field direction varies as does the direction defined by a right-handed nut advancing on a right-handed bolt, the light is right-circularly polarized, corresponding to the K^0 meson; if the direction varies as does that defined by a left-handed nut advancing on a left-handed bolt, the light is left-circularly polarized, corresponding to the $\bar{K}^0$ meson. Right-circular polarization is the sum of horizontal linear polarization and vertical linear polarization (*left*), whereas left-circular polarization is the difference (*right*). In further analogy, the wave amplitude of the K^0 meson is the sum of the K_1 and K_2 amplitudes (so that in some sense $K^0 = K_1 + K_2$), whereas the wave amplitude of the $\bar{K}^0$ meson is the difference ($\bar{K}^0 = K_1 - K_2$).

The existence of a *CP*-asymmetric force, however, could induce the observed violation of *CP* invariance. The force would have to change an antidown quark into an antistrange quark more often than it changes an antistrange quark into an antidown quark, and it would have to change a strange quark into a down quark more often than it changes a down quark into a strange quark. Since a $\bar{K}^0$ meson changes to a K^0 meson only rarely, it is possible that the transition is a chance coincidence of two moderately rare independent processes. The *CP* violation could be produced by a coincidence between a moderately improbable *CP*-symmetric transition generated by the weak force and a somewhat more improbable (by a factor of about 500) *CP*-asymmetric transition induced by a *CP*-asymmetric force. Such a moderately weak force is called milliweak (it would be roughly 1,000 times weaker than the weak force).

Clearly the set of neutral *K* mesons forms a system that provides for particularly sensitive tests of *CP*-invariance violation. Is there any other such system? Recent work, notably at the electron-positron accelerator DESY in Hamburg, has led to heightened interest in the neutral *B*-meson system, which has important parallels to the *K*-meson system. Just as neutral *K* mesons are made of down and antistrange quarks or antidown and strange quarks, so *B* mesons are made of down and antibottom quarks or antidown and bottom quarks. Bottom quarks are similar to strange quarks except they are much heavier. Hence the *B* mesons are much like *K* mesons except they are about 10 times heavier. It is my own view, however, that the greater mass of the *B* meson, among other factors, acts in such a way that we cannot expect to learn a lot about *CP*-invariance violation from studies of *B* mesons in the near future.

The milliweak-force hypothesis is attractive but suffers from the drawback that the force may not be weak enough; why has the observation of *CP* violation been limited to the neutral *K* meson? A *CP*-asymmetric force as strong as the milliweak might be expected to show up elsewhere. Could a much weaker force account for the flaw in the *CP* mirror? It would seem that this is the case. Lincoln Wolfenstein of Carnegie-Mellon University has suggested that instead of being a coincidence between a weak transition and a milliweak transition, the *CP*-invariance violation might be generated by an extreme-



CP SYMMETRY is violated by the neutral *K* meson in a second way, distinct from the preferential decay into positrons (*see illustration on page 54*). The violation involves the transition of a K_2 meson, which has a wave function that is antisymmetric with respect to the *CP* mirror, into a K_1 meson, which has a symmetric wave function. The violation is analogous to an interaction of light with a dextrose solution. The analogy begins with a beam of K^0 mesons and a beam of right-circularly polarized light. The K_1 component of the K^0 mesons decays, leaving the K_2 component; simultaneously a piece of polarizing film blocks the horizontal component of the light, allowing only the vertical component to pass through. As the K_2 mesons travel through space, a *CP* violation occurs that turns some of them into K_1 mesons (the combination is the K_L meson); analogously, a dextrose solution rotates a small component of the light from the vertically polarized state to the horizontally polarized state. Consequently K_1 mesons are generated where once there were none and horizontally polarized light is generated where once there was none. The "rotation" of the *CP* antisymmetric K_2 mesons into the *CP* symmetric K_1 mesons is currently the only known violation of *CP* symmetry.

ly weak force that induces both changes by acting only once. Labeled with the oxymoron superweak, this force, a billion times weaker than the weak force, would be very difficult to see elsewhere.

Both milliweak and superweak concepts are not so much theories as they are sets of theories. Most *CP* research has been directed toward the detection of milliweak effects. The search has turned up no positive results, but until recently the specific negative results of the experiments could not be used to exclude milliweak models, because there was no general prediction of the consequences of the models. Now a quite specific signature of the milliweak force has been put forward in theoretical research initiated by Frederick J. Gilman of Stanford University and Mark Wise of the California Institute of Technology and developed further by others. In the theory, as a consequence of milliweak *CP*-asymmetric forces, the K_2 meson not only should change partially to a K_1 meson but also should decay directly into two π mesons.

The direct decay of the K_2 meson into pairs of π mesons would differ from the decay of the K_1 meson into pairs of π mesons and should therefore be detectable. When a K_1 meson decays into pairs of π mesons, it is twice as likely to decay directly into pairs of π^+ and π^- mesons as it is into pairs of π^0 mesons. The ratio should be reversed for the K_2 meson: it should be twice as likely to decay into pairs of π^0 mesons. In the absence of direct K_2 decays the K_L meson will decay into two pions only through its K_1 part and then exhibit the two-pion charge ratio of the K_1. If there are also direct decays into two pions from the K_2 part of the K_L, the ratio will be modified. Although different theories of *CP* asymmetry predict variations, almost all lead to substantial modifications of the two-to-one ratio occurring in pure K_1 decays; the general scale of the effect should be about 30 percent.

Early experiments appear to exclude differences as large as 30 percent. Some theoretical models, however, predict much smaller effects. More recent measurements made at Brookhaven by a group from Yale University and Brookhaven led by Michael P. Schmidt, William M. Morse and me, and at the Fermi National Accelerator Laboratory by a group from the University of Chicago, the University of Wisconsin and the Saclay Nuclear Research Center led by Cronin and Bruce D. Winstein, strongly suggest that the difference is less than 10 percent. These results have served to exclude some but not all interesting milliweak models of *CP*-invariance violation. A modified experiment at Fermilab and measurements currently under way at CERN, the European laboratory for particle physics, promise to reach improved sensitivities. Indeed, the CERN group has announced preliminary results that suggest a variance of 2 percent from the two-to-one ratio occurring in pure K_1 decays. The error is sufficiently small to exclude zero. The experiment is to continue for another year in an effort to refine the understanding of the sources of error.

Within a few years the results of these experiments should show either that *CP* asymmetry is caused by a milliweak force or that the variance from the pure decays is less than 1 percent and that the *CP* asymmetry is probably caused by a superweak force. Either result will establish constraints on the form of the grand equation and bring us closer to an understanding of our universe.

THE TECHNOLOGY OF WINNING

AN ENTHUSIAST'S GUIDE TO THE 1988 WINTER OLYMPIC GAMES

Stung by competition from other nations who were using science and technology to develop better athletes and sports equipment, the United States Olympic Committee decided to help American athletes by following a similar strategy. Since 1982, USOC programs have involved scientists, engineers, and whole industries. Millions of dollars have been invested in the sports sciences, and in equipment development projects to improve the chances of U.S. athletes in international competition.

Many of the projects mentioned in the following articles were the result of these programs, and most of the authors were involved in the developments they write about. Science and technology has entered the world of sports and will be an unseen companion to competitors in the 1988 Winter Olympic Games at Calgary.

Cover photo: Olaf Soot/Black Star


300 CE

THE AERODYNAMICS OF SKIING

by Michael S. Holden, Ph.D.

When alpine ski racing began in the mid-1920s the downhill was a relatively uncharted plunge from the top of a mountain. By 1948, when a separate downhill became an Olympic event, control gates had been introduced to prevent a competitor from being exposed to unnecessary natural hazards. However, the very character of a good downhill is the speed and the danger. Its winner must display both good technique and courage in going from the top to the bottom of the mountain in the shortest time. Over the years there has been a progressive refinement of downhill technique to the point where the downhill has now been recognized as a premier technical alpine event.

In the Olympic downhill, the skier drops over 2500 feet in altitude in less than 2.5 minutes. For the downhill racer's average speed of descent, approximately 70 mph, the aerodynamic drag can vary from 18 pounds, for a skier in a good low tuck, to as much as 80 pounds for a high turning position. Comparing these forces with the typical snow-drag figure of 13 pounds for straight-running skis clearly demonstrates the importance of efficient aerodynamic performance in winning an Olympic medal. This, however, does not mean that downhill racers must strive at all times for the lowest drag position. Downhill racing at its best is a choreography of movement patterns in which there is an optimum blend of effective skiing and aerodynamics.

Wind Tunnel Simulation

During the past 7 years, we have tested over 100 skiers in the low speed wind tunnel at the Calspan Corporation. A 5 component force balance has been installed and measures lift and drag as well as the turning movements in 3 directions. It consists of a strain-gauged platform with a soft rubber mat on top. This and traction tape on the bottom of the skis prevents the skiers from being blown from the platform under the near hurricane force winds (up to 80 mph).

With three monitors, the skiers can simultaneously observe their positions from above, to the side and from the front. Computer generated lift and drag forces are also displayed on the monitors giving the skiers immediate feedback.

Blending Aerodynamics with Skiing

Prior to the 1984 Olympics, the U.S. Ski Team conducted studies to provide information on equipment and technique to aid the training of their athletes. One area which received specific attention was the blending of skiing and aerodynamics in the the downhill. Wind tunnel and on-hill studies were conducted to examine effective aerodynamic skiing positions for the various maneuvers required in running a downhill course. Straight running on smooth and bumpy terrain, long and short-radius turns, and the movements to control the destabilizing effects of rapid terrain changes were studied. In the wind tunnel, using a Data General "One" for computer simulation, times were compared for both good and poor aerodynamics for segments of a downhill course.

Close attention was given to a meaningful wind tunnel simulation of the skiing positions attained in competition. Because they stood on a force balance in a tunnel where the air flowed perpendicular to gravity, athletes—left on their own—were found to adopt positions with their backs horizontal. However, during competition, in all but the low tuck, racers actually ski with their thighs approximately parallel to the slope and with their backs rounded at between 20 to 60 degrees to the slope—a position which is not naturally held in the tunnel. The arm positions that are most effective in minimizing drag with the back flat differ significantly from realistic positions with the back inclined. Thus, a poorly conceived wind tunnel study can lead to totally misleading suggestions in training. Only by alternating studies on the slopes and in the wind tunnel can a meaningful training program be conducted.

Aerodynamic Skiing Forms

The aerodynamic drag of a skier is composed principally of form drag, which results from the pressure difference across the body as flow separation occurs over the rear of the body, and skin-friction drag. For a downhill racer, the skin friction drag forms only about 2% of the total drag. To lower the drag of the body, the racer attempts to attain a streamlined shape and minimize the area over which pressure difference acts. Figure 1 shows a typical "low" tuck position developed in wind tunnel studies with members of the U.S. Olympic team. By holding each part of the body to maximize its aspect ratio (length in the flow direction divided by length perpendicular to the flow), the shape is more streamlined and the projected area minimized. Lengthening of the body in the flow direction also allows the transition from laminar to turbulent flow to occur on the body, thereby decreasing the size of the wake and increasing the pressures over the back of the body. The drag coefficient of a typical low tuck is 0.28, a figure comparable to a streamlined automobile. A major mistake made by many ski racers is to employ the arm positions used in a low tuck when skiing in more upright body positions. For skiing on other than smooth, straight segments

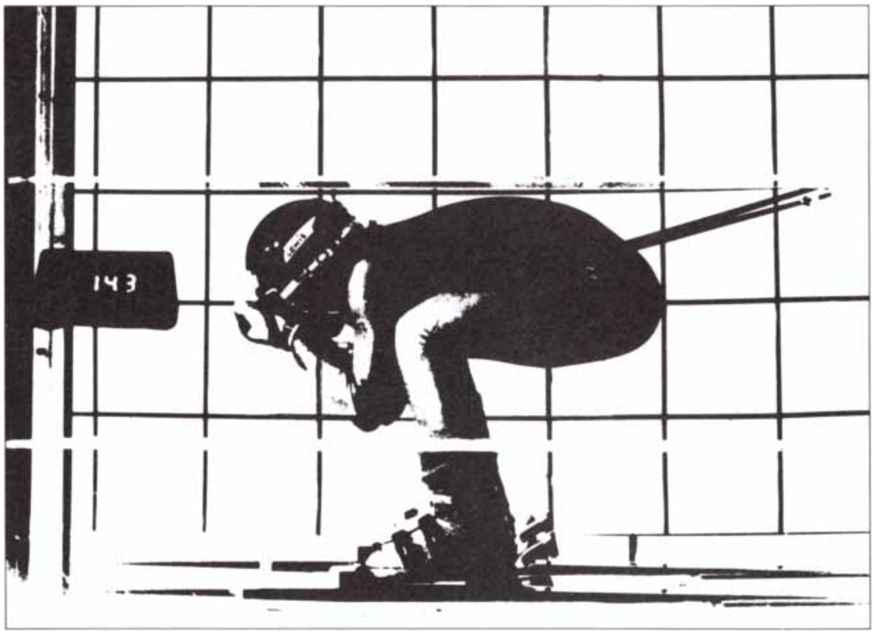
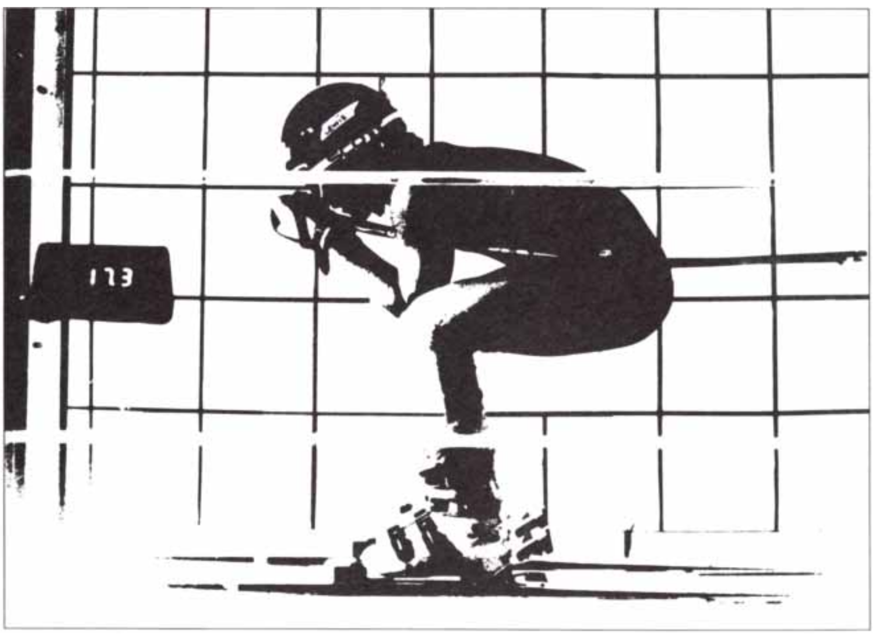


FIGURE 1 The low tuck is the skiing position with the lowest drag and is employed in smooth straight sections of the course. The body is held with chest almost horizontal, the arms are aligned with the lower legs and the hands held close to the face.



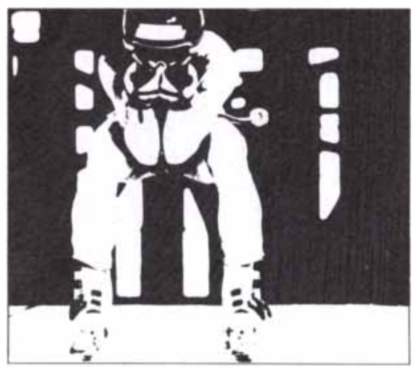


FIGURE 2 In a good "gliding" tuck, which is employed during long radius turns or in bumpy terrain, the elbows are pulled together in front of the chest to round and lengthen the upper body. Alternatively, with a small increase in drag, the hands can be lowered and pushed ahead, with the elbows close, to divert the air around the chest.

of the course, the racer adopts body positions with the thighs approximately parallel to the slope and the back at an angle to it. For such skiing positions, the lowest drag is obtained with the elbows pulled together and the forearms held close to the chest as shown in Figure 2. This rounds out the front of the upper body and generates a larger aspect ratio. Alternately, with only a small increase in drag, the forearm can be lowered and pushed ahead to divert the wind around the chest. Many racers employ low-tuck arm positions with an open chest, resulting in a relative drag increase and ineffective skiing positions.

Recently, bumps or jumps have been added to a number of downhill courses to add more "excitement" as the skier is launched and travels many feet in the air. Controlling the take-off and flight and absorbing the landing are important technical skills. To lower initial flight trajectory, a "pre-jump" is employed where the racer rapidly drops his body just before take-off. Employing the correct movement patterns during


GATE

take-off, flight and landing can prevent loss of time and control. The adage that it is faster staying on the ground than taking to the air is true only because the body extension necessary to control the take-off and landing creates greater aerodynamic drag.

Ski Jumping

In addition to wind tunnel testing downhill racers, we have also built a special test rig to train ski jumpers. The jumper is supported in the tunnel by three discreet systems, a harness to the upper body, a harness to the thighs and a pedestal to which the skis are attached. The pedestal in turn is supported by the force platform described previously. The soft suspension system allows skiers to fly in the wind tunnel with three feet of vertical and one foot of horizontal travel. The test rig is quite sophisticated with gear motors to position skiers in the wind tunnel, a torsion beam suspension system, and a display of lift and drag on the monitors. The skiers can be placed in realistic positions with respect to the wind stream, and they are able to try body and ski orientations that they have observed or that they think may improve performance.

Even though the feel of the system is not entirely realistic, due to the fact gravity is displaced by about 30 degrees from that experienced by a jumper in flight, the information generated is extremely valuable. A computer simulation gives each jumper their estimated distance as they experiment with positions corresponding to various phases of the jump.

In order to maximize distance, a ski jumper must initially use an angle of attack that produces a maximum ratio of lift to drag for efficient gliding. As the flight progresses, the angle of attack must gradually become steeper, producing maximum lift and correspondingly high drag. In this position, the skier flies in a complete stall, but the high lift and high drag both have components perpendicular to the slope, and help increase the distance. The problem is trying to train skiers to properly orient themselves without the horizon for reference. A possible but expensive solution would be to build a special wind tunnel tilted at 30 degrees so that the gravitational vector would be properly oriented.

In spite of some problems, the downhill and ski jump training systems have provided an excellent learning environment for the skiers. The visuals give immediate feedback, so that the skier can determine which individual positions work best. All of the current U.S. downhill team, and most of the jumpers have been trained in the Calspan wind tunnel.

Gliding

In combination with good aerodynamics, gliding well on skis requires special techniques to control ski-snow interaction. While the type of skis and their base preparation control the friction generated between the skis and the snow, there is a component of the force in the direction of motion that is actively generated by the skier (see "The Technology of Winning Alpine Ski Races"). Even when running straight downhill, active foot control of the ski causes "micro" turns, which generate significant "induced" drag forces at high speeds. The good "gliders" are those who, by subtle foot control, minimize this induced drag. We have begun field studies to measure the actual forces between ski and boot so that these forces may be understood more clearly.

Skin-Tight Suits

It was not until the late 1960's that one-piece stretch suits came into extensive use by downhillers to reduce air resistance. Olympic regulations specify that all articles of clothing worn by the skier must conform to the general contours of the body, thereby precluding the use of aerodynamic fairings. The first stretch downhill suits had an outside coating of shiny neoprene, a construction still used in many other Olympic sports. However, a spill on a downhill course in one of these suits could result in the skier picking up rather than losing speed, when traveling in a prone position. Although this problem was resolved by moving the coating to the inside of the suit, medical problems associated with the body's inability to "breathe" in such a suit have resulted in regulations mandating a minimum material porosity. Interestingly, suit porosity has a significant effect on drag, even though suit surface roughness is insufficient to increase skin friction drag. A racer clad in a porous suit will, in general, have a drag 5 percent greater than when wearing an externally or internally neoprene-sealed suit. It has been speculated that this increase in drag results from macroscopic flapping of the surface, or perhaps an adverse change in the position of separation, resulting from flow through the suit. Also, ventilation through the suit can trap more air thus increasing the apparent mass of the skier and causing a higher drag due to momentum transfer. Devices to trip the boundary layer prematurely from laminar to turbulent, thereby filling in the low pressure void behind the skier, have been tried without success.

During the past ten years, there have been significant changes in the technique of downhill racing, the most important of which is the combining of effective aerodynamics and good skiing. Developments in ski design, combined with smoother and more consistent course grooming, have resulted in the ability of top competitors to ski with greater and more subtle control, allowing them to adopt and hold effective aerodynamic body positions. The introductions of the skin-tight racing suit and improved ski base preparation techniques have resulted in significantly increased skier speeds.

While traditionalists may mourn the passing of the arm-flailing performance of Franz Klammer in the 1970 Olympics, the knowledgeable spectator will appreciate the contemporary, more refined yet aggressive techniques employed to conquer the mountain.

Michael S. Holden, Ph.D.
Principal Scientist, Calspan Corporation. Extensive research in hypersonic aircraft including the new spacecraft NASP-X30. Expert in aerodynamics of downhill racers and ski jumpers, has conducted U.S. Ski Team wind tunnel training since 1980.

THE AERODYNAMICS OF THE BOBSLED

by Paul Van Valkenburgh

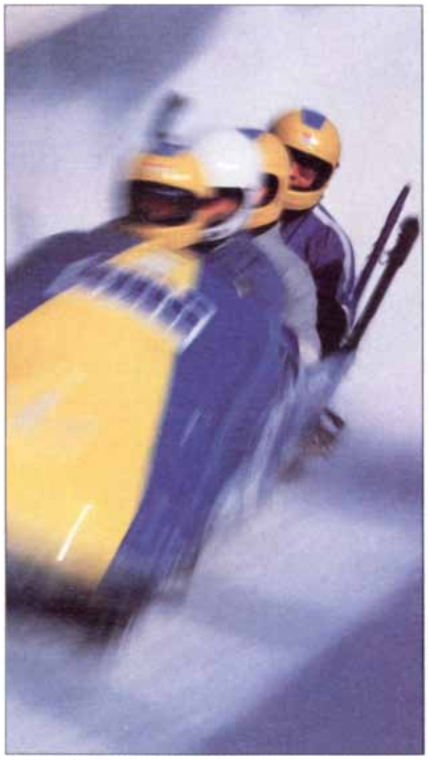
A researcher doesn't have to go too far back in history to review the primitive beginnings of bobsled technology. A mere decade ago, aerodynamics, runners, and chassis structures were all designed intuitively—based on years of experience by a handful of European craftsmen. Even today, bobsleds have essentially no suspension and are steered by ropes connected through a linkage to a wagon-type front axle.

Bobsled runs use refrigerated ice, and are from 1200 to 1600 meters long with vertical drops of 110 to 150 meters. The sleds reach speeds over 80 mph while careening around hairpin turns and pinning the riders down with 4 to 5 G accelerations. To drive the courses successfully requires lightning reflexes, enormous skill and intensive practice.

American attempts at scientific 2-man bobsled design have had mixed results. A 2-man design by Dale Smith placed 5th in the 1980 Olympics, but many other projects since have produced unstable sleds that were not used in competition. In 1984, Bobsled Federation President William Napier asked Dr. Robert Gilmore of General Electric to fill this challenging technological vacuum.

The key areas of research included mathematical simulations to determine the relative value of potential modifications, aerodynamic wind tunnel and math modeling studies, runner friction measurements, and on-board instrumentation for ride, handling and push start research.

Chassis strength, stiffness, weight, and sophisticated suspensions were not major considerations because a generous weight allowance makes almost any chassis design feasible. The present primitive suspension is mandated by the regulations.


E.M. Bordis/Leo de Wys Inc.

Computer Modeling

Computer modeling has been undertaken by a number of researchers. However, as usual the greatest obstacle is not the mathematics, but the measurement or estimation of all the relevant parameters. A fairly good measurement of the courses—length, elevation, and banking—is available from construction drawings and surveys, although the precise path of the sled is not so well defined. Aerodynamic drag of the sleds has been estimated fairly accurately from wind tunnel tests. However, ice friction coefficients are still only a rough approximation, especially when considering the contributions of side-slip and vibration.

A simple computer model which included push-start energy input and best estimates of aerodynamic and ice-runner losses, showed that the ultimate margin for improvement is about 3 seconds in a total run time of 59 seconds; the sleds are approximately 95% efficient. To compare the relative value of aerodynamics and runner friction, which have considerably different influences as a function of speed, Airflow Sciences of Livonio, Michigan used an energy balance equation. They estimate that the average relative losses to air and mechanical friction during a run are 20% for air versus 40% for ice for the 4-man sled, and 26% versus 38% for the 2-man sled. Although runner friction is smaller than aerodynamic drag at higher speeds, at low speeds the reverse is true; because proportionately more time is spent at lower speeds runner friction dominates overall.

Sensitivity studies were also made of the effect of sled weight and push-start performance. As might be expected on a coasting vehicle, not ballasting to the maximum weight allowance (630 kilos for the 4-man, and 390 kilos for the 2-man), costs some time, or about 0.3 extra seconds for being 5% underweight. Early force balance computer models by a team of engineering students under Professor Robert W. Manning at Syracuse University and work by Airflow Sciences, indicated the critical importance of the push start, predicting that every 0.1 second saved at the start translated into 0.3 seconds at the finish. Dr. Jim Richards at the University of Delaware has been using a sled with strain-gauged push bars to evaluate individuals and improve their start performance. As a result, the decision was made to select a crew from very quick sprinters who could provide a higher top push speed, rather than weightlifters who might be faster off the line.

Ice-Runner Friction

Even though ice-runner friction would seem to be a fruitful area for improvement, after a great deal of research it is still not possible to accurately predict the friction coefficients of runners on ice. This is not surprising since bobsleds not only slide, but frequently gouge their way down the run leaving deep tracks in the ice surface. The science which describes this, tribology, is one of the most difficult in modern research. Measurements have shown that ice-runner friction can range from about 10% to as low as 0.5% of the normal force on the ice depending upon the slope, speed, runner shape, ice temperature, ice impurities, and a large number of other difficult to control variables.

Experiments at Dartmouth College and at the Cold Regions Research and Engineering Laboratories, using cylindrical sled runner samples sliding on a rotating cylindrical block of ice in a strain-gauged lathe feed, have been most relevant. Although the contact geometry is more localized than a bobsled runner, it is at least constant and the variables are easy to control. The experimental properties that were isolated and analyzed included ice temperature and roughness, the temperature, roughness, hardness, and thermal conductivity of the runner, and sliding speed. While smooth surface hard steels appear to have lower friction coefficients at -5 degrees C., a softer alloy with a lower thermal conductivity appears to work better at -10 degrees C. However, overall hard steels are best because they will stay smooth longer and withstand scoring from abrasive impurities in the ice.

The average ice friction for a typical run has now been narrowed down to around 1.5 to 5% of the total weight on the runners. Pressure and speed can actually melt the ice and cause the runner to glide on a microscopic film of water. Currently, Airflow Sciences is making an attempt to measure the low speed drag of real runners by the "coastdown method." A weighted sled coasts along a 16-foot ice track on a tiltable truss, with timing lights to measure the change in velocity. The friction force can be calculated from the speed-time-distance behavior.

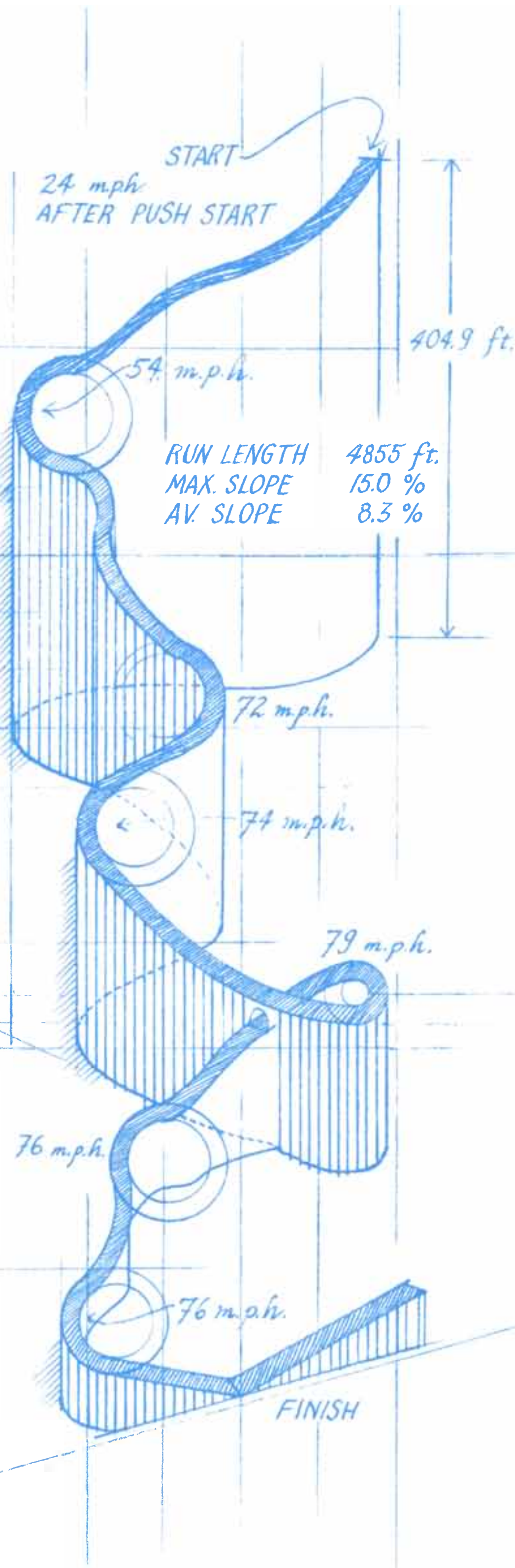
Aerodynamic Drag

Aerodynamic drag, even with its less important contribution, seems to be the most productive area of research. A body shape is harder to define specifically in the regulations, and therefore leaves more opportunity for optimization. In 1985, Airflow Sciences concentrated on development of a 2-man sled in wind tunnels at National Research of Canada and the University of Michigan while the Flight Research Institute of Seattle took a 4-man sled to the University of Washington tunnel. Baseline studies showed the best 1984 sleds to have drag coefficients of about 0.45 with equivalent drag areas of 2.0 sq. ft. for the 2-man and 2.5 sq. ft. for the 4-man (drag area is the projected frontal area times the drag coefficient).

Initially, it was decided to work on sleds of known stability and performance. Thus 1985-86 efforts were focused on aerodynamic modification of existing sleds. Air drag was reduced by about 16% for the 2-man sled, and 11% for the 4-man, by straightforward changes in rider position and windscreen angles, and by making fairings for the runners and push bars. With these improvements, the energy balance computer model showed a run time reduction of almost a second, equivalent to a typical spread between first and eighth place.

In 1986-87, work on the U.S. sleds produced an extra 20% drag reduction. But a question of rules interpretation prevented these sleds and those of many other countries from running. The sleds were rapidly modified to meet the rules and in the 1986-87 World Cup the U.S. Team placed first in the 4-man, second in 2-man and first in combined.

As a result of the rules difficulties, Dr. James Paul of Airflow Sciences was appointed U.S. liaison to the International Bobsled Federation Rules Committee. He met officials and other builders in Switzerland to clarify the rules, and then everyone went back to their wind tunnels. Again, most drag reductions came from changing the position of the riders behind the driver



The three-dimensional sketch of the bobsled run at Calgary illustrating the curves, grades and speeds. The final uphill section is the run-out area where the brakes are applied after the finish.

Again.

Cherokee 2-Door.

Jeep® Cherokee
4-Wheel & Off-Road 4x4 of the Year

Jeep Eagle

to give a lower profile, but there were also a large number of improvements in small areas, such as the underbody, the runner-body interface, and the blunt "boat-tail," which must be left open for brakeman entry.

New Sleds

Ultimately, the wind tunnel models produced drag reductions of over 40% from the '84 baselines. At that point, full size clay models of both 2 and 4-man sleds were made and turned over to EG Composites, an Indianapolis race car fabricator, for construction in carbon fiber, Kevlar and titanium. Along the way, the decision was made to utilize race car monocoque construction. This was primarily for safety—in terms of impact strength and elimination of internal constructions—but a finite element analysis produced a weight reduction of about 200 pounds, which could be used as moveable ballast to improve handling.

This revolutionary chassis raised the question of driver-acceptability of the unanticipated vehicle dynamics effects. So the team rented the Winterberg, Germany bobsled run for private tests to check out the handling and familiarize the crew. As a backup, Jerry Baer at the Flight Research Institute continued to make evolutionary changes in aerodynamics on the conventional 4-man bobsled frame, with 15-25% drag reductions, primarily from moving the driver forward, and allowing the others to assume a lower profile. Cowl changes and streamlining the driver's helmet also improved the drag.

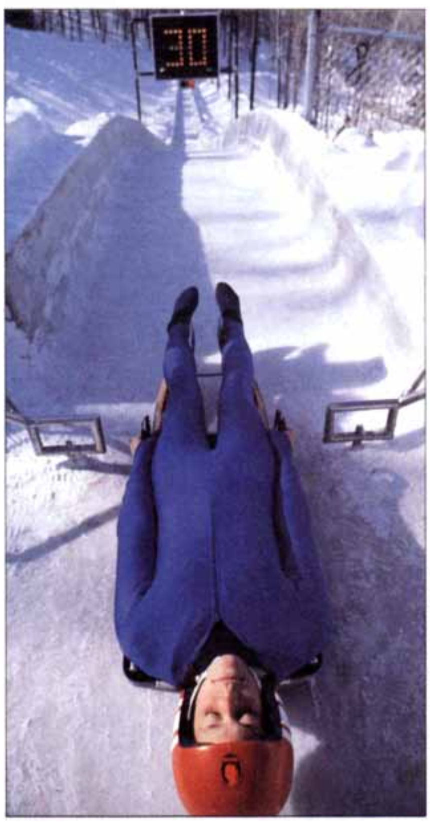
A backup evolutionary 2-man sled was designed and fabricated by a group of students at Syracuse University led by Professors Jerome Malinowski and Robert Manning. All of the sleds with the exception of the Airflow Sciences 4-man were tested at Winterberg.

Instrumentation

The ultimate measure of all these efforts is elapsed time on the actual course. But, this is so dependent on driver variability—especially with the learning required in modified sleds—that only on-board data recording can sort out the effects. A commercial 2-channel digital system was used to measure speed and G's and a 4-channel analog recorder was used for high frequency vibration inputs.

Transients and vibration are of interest not only for structural analysis, but to try and understand the ice roughness effects on stability and control. Historical observation of the pounding the sled and runners receive was substantiated by accelerometer recordings showing almost random vibrations at many G's. It turns out to be analogous to driving a car on concrete with steel wheels and no suspension.

Continuous speed recordings may seem irrelevant compared to the goal of overall elapsed time, but not to the "drivers-in-training." As had been seen with race car drivers, an initial reluctance or outright apathy on the part of the test drivers soon gives way to excitement as they can actually measure—corner by corner—where individual differences in performance lie.



Michael Norcia/Leo de Wys Inc.

Bobsledding is an exciting new technological challenger, limited only by a few rules, a short season, a small number of tracks, and a limited R&D budget. But it looks as if American innovation and technology has again made a significant contribution to athletic endeavors.

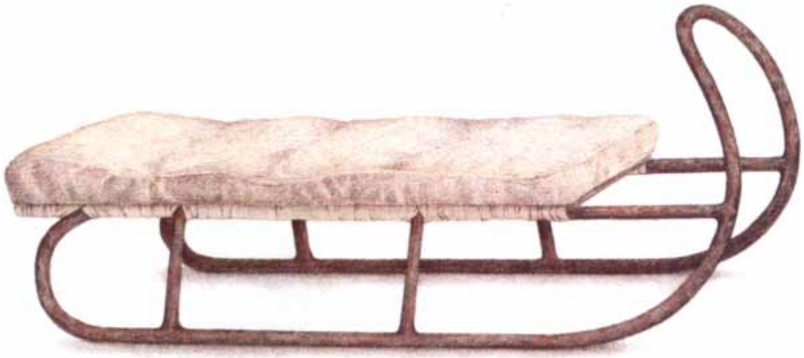
Paul VanValkenburgh
Aerodynamicist, race car engineer, author and consultant. Specialist in automotive design, testing and analytical instrumentation. Designed test instrument system for the U.S. Bobsled Team.

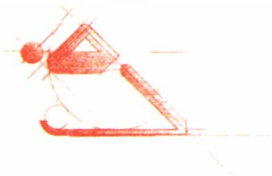
NEW SLED FOR THE LUGE

Luge sleds are ridden at terrifying speed through a serpentine ice run, by athletes lying on their backs with feet forward. A luger hangs on for dear life while steering with foot levers attached to the runners. The sled is carefully formed to the human shape and is streamlined in every aspect. A luger's dress includes gloves, a safety helmet, a face shield and slick rubberized clothing that clings perfectly to the human body.

The top U.S. woman luger, Bonnie Warner, who was third overall in the 1986-87 World Cup competition, is normally within 0.3 seconds of the West Germans who are currently the best in the world. She thinks she can improve 0.1 second in the start, which would still leave her 0.2 seconds shy of winning at Calgary. Bart Hibbs of AeroVironment Inc. is working to improve the aerodynamics of Bonnie's sled and clothing. He has calculated that a 9% air drag reduction will gain the missing 0.2 seconds.

Wind-tunnel tests show that fairing the hands into the sled, along with other revisions, lowers air drag by 10%. Hibbs is also planning to put riblets on the sled and on Bonnie's clothing. Riblets, manufactured by the 3M Corporation, are parallel microscopic grooves in the surface of a plastic adhesive film. They gained notoriety when used successfully on Dennis Conner's America's Cup yacht. If Hibbs' calculations are correct the riblets should further reduce air drag by 2%.





THE SKI JUMP, FLYING WITHOUT WINGS

by Chester R. Kyle, Ph.D.

Leo de Wys Inc.

John P. Kelly / The Image Bank

Picture yourself on a narrow bar at the top of an Olympic ski jump so nearly vertical that it resembles the first drop of a rollercoaster. You are more than 300 feet above the landing area below. The hill falls steeply from there to the flat out-run area where you must stop. There is an elevation difference of 450 feet from the starting bar to the out-run.

You are dressed in a safety helmet and a tight one-piece suit. Your skis are huge, over 8 feet long and 4½ inches wide. Your boots and cable bindings allow your heels to lift off the skis so that they are nearly parallel to your body forming a wing-like lifting surface. You push off from the bar and shoot down the in-run track tucked in a low crouch with your arms flattened against your sides. There is no turning back.

You accelerate rapidly to 60 mph before hitting the take-off ramp, and with your weight forward you strongly thrust up and outward into space. Quickly you lift the tips of your skis upward and change to a graceful gliding position with your body rigid and arms at your sides. For just over three seconds, you are in free flight floating downward almost parallel to the hill. As the snow rushes upward, you glide till the last possible instant, and then, pulling your skis under you and with one leg forward in telemark position, you absorb the shock of landing.

For an instant your body weighs four times what it does ordinarily, but you survive the landing and flash through the transition zone to the flat out-run area below, skidding sideways to a stop. The huge crowd roars its approval. It is obvious why ski jumping is one of the most thrilling of the Winter Olympic events.

John P. Kelly/The Image Bank

Errors in take-off or in flight, or quirky winds, can cause a jumper to turn completely over and crash. A bad landing can send a skier tumbling down the slope in a frightening tangle. How could anyone have the nerve or skill to try such a potentially dangerous sport?

The commonly held impression that ski jumping is dangerous is misleading according to Tom Daggett, Technical Director of the U.S. Ski Jump Team. Daggett points out that a University of Missouri study found ski jumping at the elite level to be safer than cross-country skiing, and 8 times safer than downhill skiing.

The distances traveled in the air are phenomenal. In the 70 meter jump they cover up to 300 feet, and in the 90 meter, nearly 400 feet.

Actually there is a danger in jumping beyond the critical point where the transition curve begins. If a jumper goes too far in competition, the judges have the right to move the starting bar down the slope to lower the take-off speed. If the start is relocated, all jumps during that round are cancelled and the round is repeated. Each jumper is allowed two jumps upon which they are judged. They are awarded 60 points for distance if they come down midway in the landing area. Points are added or subtracted for being long or short of this mark. The judges also award up to 60 points for style and form, so not only must competitors jump long distances, they must do it with grace and precision.

Distance is primarily a function of two factors, take-off velocity and take-off technique. Velocity depends upon the length of the in-run which is carefully controlled by the judges. When this is fixed, jumpers try to maximize their speed by using the most efficient aerodynamic posture possible during the in-run. Individual wind tunnel studies help each jumper to determine this.

On the take-off, the jumper must choose the best trajectory for maximum distance. This requires split second timing and coordination. If they lift too high, it kills speed and they fall short. If they dive too low they also fall short. Taking off from the very end of the ramp, using exactly the right lift, changing rapidly to the most efficient glide posture, and flying at the best angle of attack all the way down, produces the optimum glide path and the longest jumps. The key words are quick and quiet, with no unnecessary movement to disturb airflow.

Efficient aerodynamics is critical to the ski jump. Earlier jumpers used bad posture on the in-run, and windmilled their arms in flight. They also wore sloppy clothing, and used narrow skis. All of these produced high air drag and less lift, making jumps much shorter than today.

Present ski and clothing design is very efficient but it is carefully controlled by the rules. At one time the Austrians used balloon suits allowing air to enter the front, trapping it in back and forming the skier's body into a nice airfoil shape. These suits were outlawed when the Austrians began out-jumping everyone. The Austrians also put a flap on the tail of the skis, to produce a higher lift. This was quickly banned. Helmet form is also specified to prevent outlandish airfoil shapes from being used.

Even with rigid rules there is still latitude for improvement. Mont Hubbard of the University of California at Davis is modeling the flight of a jumper to predict the optimum glide path for each skier. Body shape and weight vary, so this optimum path would depend upon the individual, (the best jumpers at present seem to be very light and thin, almost bird-like). Better waxes and low-friction skis can give higher take-off speeds. Better clothing, boots, bindings and skis can still be developed within the rules, and better training techniques can be devised.

During an entire career, a ski jumper only spends a few minutes in the air. Special practice wind tunnels could be built to give jumpers several hours to refine their flying skills. Above all, as more jumps are built, better athletes will be attracted to the sport. With this, the techniques will change and improve, and this exotic sport will no doubt be more exciting than ever.

Chester R. Kyle, Ph. D.
Sports equipment consultant. Designer 1984 Olympic bicycles and clothing for the U.S. Cycling Team. Member U.S. Olympic Sports Equipment and Technology Committee. Founder and past president, International Human Powered Vehicle Association. Science Editor, Bicycling Magazine.

BUICK

NORDIC SKI RACING

Biomedical and Technical Improvements in Cross-Country Skiing

by E. C. "Ned" Frederick, Ph.D. and Glenn M. Street

The nordic skiing sports of biathlon, cross-country skiing, and nordic combined all have as their main ingredient skiing over rugged terrain while wearing narrow cross-country skis and using lightweight poles to aid propulsion. Cross-country skiers simply ski. However, athletes in biathlon and nordic combined must develop other skills. Biathletes are required to carry a .22 rifle and stop during the race to knock down targets—skiing penalty loops or suffering time penalties for the targets they miss. Nordic combined competitors are handicapped in their ski race based on the results of their ski jumping competition held the day before.

Although the unique requirements of each of these nordic events elicit subtle refinements in the ski technique and the equipment used by the athletes, all nordic skiers have two main options for getting around. They can use the classical diagonal stride technique, or the new and swifter freestyle ski skating technique.

E.M. Bordis/Leo de Wys Inc.

Classical Skiing versus Freestyle Skating

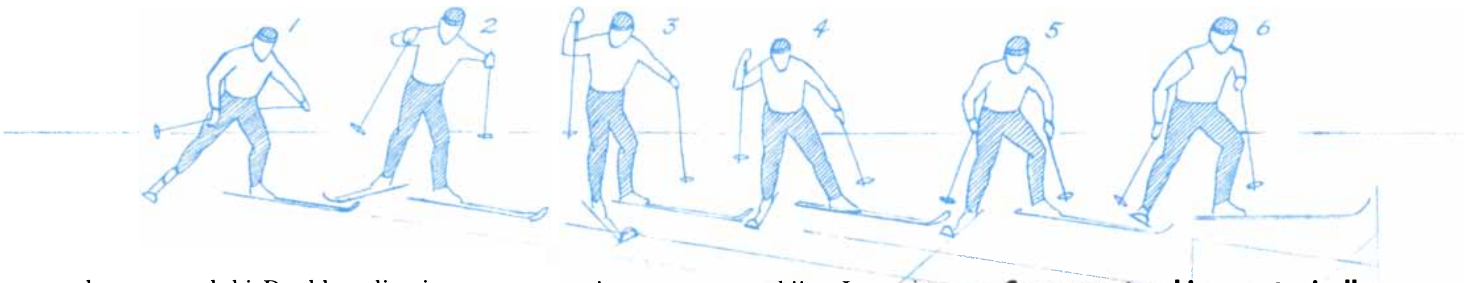
Diagonal striding involves skiing in parallel tracks which are carved into the snow by special track-setting machines. The friction required for forward propulsion is produced by what is called the "kick". In the kick the skier pushes the "kick zone" of the ski base just under the foot into contact with the snow. The tacky kick wax applied to the kick zone produces enough traction to allow the skier to stride forward and upward of the kicking ski and onto the "gliding ski". This striding motion is aided by the alternate use of ski poles which apply a carefully timed horizontal impulse to drive the skier forward.

Diagonal striding was the dominant technique used by nordic skiers until the early 1980's when American Bill Koch won the cross-country skiing World Cup using predominantly skating techniques. Since then skating has dominated cross-country skiing and has stimulated a number of innovations in skiing equipment and technique.

The term "skating" is descriptive of the swaying push and glide motion reminiscent of speed skating. Skating offers a number of biomechanical advantages to skiers. For one thing, skaters don't use slow kick wax. They can use stiffer skis with glide wax alone. This means better glide especially on the downhill sections of courses.

Skaters get their "kick" by gliding out to the side at an angle of between 10 to 20° to the direction of travel, and tilting the inside edge of the ski into the snow to push off. The fact that the ski interlocks with the snow and that it continues to glide as the skier pushes off means that kicking forces are more effective. Also, more of the upper body musculature is used since the skater double poles, planting both poles at once; the arms and massive trunk muscles as well as the legs are used for propulsion. In diagonal striding, the skier poles on alternate sides with a more erect position.

Skating has other advantages as well. Body weight is transferred from ski to ski as in the diagonal stride, but in skating the body falls forward onto the gliding ski. In the diagonal stride the skier jumps onto the other ski, lifting the mass of the body in each step before landing on the gliding ski. This leaping stride is metabolically more costly than skating's gravity aided weight transfer. Another energetically sensible advantage of skating is the fact that velocity changes of the body's mass are less extreme—the skier pushes off a moving rather



Cross-country ski racers typically use the V1 pattern where double poling occurs on one side only. Note the pronounced flexion and extension of the trunk during this half cycle. In thisstyle of skating, these movements are repeated on the opposite side to complete the full cycle. (Distance is not to scale.)

than stopped ski. Double poling in skating can be timed to further minimize the detrimental bleeding off of speed in each glide as well.

These factors make skating faster. How much faster? One of us (Street) has shown that, on average, Olympic caliber skiers are winning races about 18% faster than when only classical techniques were used.

Because skating is so much faster, biathletes and nordic combined skiers use it exclusively in their competitions. In cross-country skiing, officials have mandated that half of all championship races be classical (skating restricted) and half freestyle (skating unrestricted). For this reason, the Olympic program in Calgary will include both. The shorter races of 15 and 30 km for men, and 5 and 10 km for women will use the classic technique (however even here skating is permitted in certain designated areas). The longer races, 50 km for men and 20 km for women plus all of the relays as well as the biathlon and nordic combined events will use the freestyle skating technique.

All cross-country skiers must overcome the physical constraints of gravity and snow and air drag which are the consequences of trying to move about on snowy hilly terrain. Until zero G space skiing becomes the rage, we are stuck with gravity, but there are things that can be done about snow and air drag. In fact, most advances in cross-country skiing are the result of innovations that either reduce these resistive forces, or enhance the effectiveness of the biomechanical forces generated to overcome them.

Snow Drag

Snow drag is the resistance that snow presents to a gliding ski, and it is the most important factor limiting performance in cross-country skiing. It acts in opposition to the gliding ski, and viewed simply, it is the product of the normal force and the coefficient of snow drag. Normal force is applied perpendicular to the surface under the ski, and it is the result of both body weight and forceful actions of the legs. Normal force fluctuates during each ski stride between zero and about 3 times body weight.

The coefficient of snow drag is the result of two principal factors, the friction between the sliding ski and the snow and the combined compression and deflection of the snow, the "packing and plowing" action. The coefficient of snow drag varies between about 0.05 and 0.2 depending on the snow character, the temperature, and ski preparation and design.

Skis, Boots, Poles and Waxes for Skating

In design, skis that glide better on softer snow are generally not as stiff and have a more even pressure profile. Skis that glide well on hard snow are stiffer and distribute relatively more pressure to the tips and tails than to the mid-section. Skating skis, for instance, generally distribute pressure more evenly than classic skis, they are about 5 cm shorter, and have a higher torsional stiffness to aid in applying kicking force. Although rules prohibit metal edges, some skating ski edges are reinforced with steel under a thin layer of polyethylene, making them harder and sharper than classic ski edges. Other factors such as tip and tail shape also significantly affect glide performance.

Glide performance can also be enhanced by other equipment modifications. Although there has been little change in competition cross-country ski bindings, the boots are stiffer laterally and provide more control. Poles have also changed. They are stiffer and about 20 cm longer. The use of high strength aluminum and carbon fiber has also decreased the weight of the longer poles.

Skis, waxes and base preparations are designed to reduce snow drag. Improvements in the polyethylenes used in covering ski bases occur regularly. The introduction of black carbon impregnated bases in recent years has resulted in improved glide performance on dirty, old snow. Several innovations in wax formulation have also resulted in better glide performance. The use of graphite powders, silicones, even Teflon preparations have both broadened the range of certain waxes and resulted in products that better deal with difficult snow conditions.

If we split the factors that affect snow drag into ski-design-related and ski-preparation-related variations it becomes apparent that ski design is the more critical factor.

A comparison of controlled glide tests on 13 models of skis provided by 6 different manufacturers and all designed for ski skating showed a range of 8% in average elapsed time through two timing gates.

When notched skis are used with a range of base preparations and with a variety of waxes designated for the conditions, the variation in glide test results is about 5%. The knowledge and skill of the ski tuners who prepare and test the skis used in competition operates within that 5% range. Taken together, ski design and ski preparation then might account for as much as 13% of glide performance.

We hasten to point out, however, that gliding is not the only component of forward progression and so glide performance cannot be equated to overall performance. In fact, about 20% of the motion in a ski race comes from forward displacements of the body's center of mass, such as stepping forward, produced independent of glide. This means that if we want to get a more accurate picture of the overall effect of the ski and snow drag on the performance of Olympic caliber cross-country skiers, we should reduce the 13% glide variation by 1/5th to just over 10%. Therefore proper ski design and preparation can mean about 8 minutes in a 30km cross-country race.

Aerodynamics

Aerodynamics can provide another area of improvement for nordic skiers. Elite skiers can skate a 30 km course at an average speed of 14.5 mph. At this speed only about 10% of the frictional force is air resistance. However between 35 and 45% of the distance in a typical race is on downgrades where speeds can be from 20 mph to over 30 mph. Here aerodynamic drag is the major retarding force.

Elite skiers are beginning to use proven techniques to decrease air drag. Their uniforms are made of smooth tight Lycra, similar to speed skating suits. Wind tunnel tests have shown them to have from 6 to 10% less drag than traditional wool suits and oversize caps. Lower profile ski tips and streamlined poles also lower wind drag.

An egg shape tuck position can decrease wind resistance as much as 50%. However cross-country skiers often use downhills to rest and a crouched descent position is almost too much to ask of rubberized thigh muscles. Nevertheless, by using known aerodynamic drag-reducing techniques, race times could be from 2 to 3% faster. At present, most coaches and athletes do little to improve this aspect of performance.

Winning cross-country ski races means going faster than your competitiors. Skiers who want to go fast skate, and they reduce snow drag by using ski designs, waxes and base preparations that minimize deflection of snow. These factors have the potential to enhance performance by as much as 10%. Further improvements, on the order of 3%, can be made by streamlining clothing and equipment and by refining aerodynamic postures on downgrades.

Kevin Sulgert/Index Stock Inc.

E. C. "Ned" Frederick, Ph.D.
President Exeter Research, specializing in sports biomechanics. Advisor in biomechanics to numerous Olympic athletes and to U.S. Biathlon Team. Author, over 100 scientific articles; active Biathlon competitor. Former head of Nike Research.

Glenn M. Street
Head of U.S. cross-country ski equipment development project. Consultant to U.S. Ski Team. Active ski racer. Finishing Ph.D. studies at Pennsylvania State University. Assistant Professor at St. Cloud University.

OXYGEN CONSUMPTION OF NORDIC SKIERS

Studies show that the ability of cross-country skiers to absorb oxygen into their blood is higher than any other Olympic athletes including marathon runners. Nordic skiers have a maximum of over 80 milliliters of oxygen uptake per kilogram of body weight per minute whereas marathoners average about 75. For comparison, middle distance runners, rowers, and swimmers average about 62, while soccer players average less than 50. This means that nordic skiers are probably the best endurance athletes in the world.


Steve Brown/Leo de Wys Inc.

SHARPENING FIGURE SKATES

Engineer Sidney Broadbent has carefully measured the characteristics of ice skates used by elite skaters and concludes that they are often inaccurately sharpened which could result in poor performance. If the hollow ground edge is not centered properly, or if the edge is rough and uneven the skater's control could be highly affected. Broadbent has designed an automatic machine to inspect the accuracy of skate blades, and will advise elite skaters on whether more precise sharpening is required.

For Bill Demby, the difference means getting another shot.

When Bill Demby was in Vietnam, he used to dream of coming home and playing a little basketball with the guys.

A dream that all but died when he lost both his legs to a Viet Cong rocket.

But then, a group of researchers discovered that a remarkable Du Pont plastic could help make artificial limbs that were more resilient, more flexible, more like life itself.

Thanks to these efforts, Bill Demby is back. And some say, he hasn't lost a step.

At Du Pont, we make the things that make a difference.

Better things for better living. DU PONT

REG. U.S. PAT & TM OFF

John P. Kelly/The Image Bank

SCIENTIFIC AMERICAN's commitment to The Technology of Winning continues in the September, 1988 issue—with a focus on the Summer Olympic Games.

SCIENTIFIC AMERICAN and the following sponsors of "The Technology of Winning" are making additional information available, free. Circle the number beside each item you wish to receive, clip the form and mail to:
SCIENTIFIC AMERICAN – February 1988 P.O. Box 338 Dalton, MA 01227-9963

1 The American Express® Card. "Membership Has Its Privileges℠."
To apply, circle #1, or call 1-800-THE-CARD.

2 Buick Motor Division

3 NYNEX Corporation

4 20 minutes every other day to total fitness. Only the SCHWINN AIR-DYNE exercises both your upper and lower body simultaneously.

5 "The Aerodynamics of Skiing" – Graphic depiction of aerodynamic skiing forms with further elaboration by the author.

GABAergic Neurons

Nerve cells not only excite their neighbors but also inhibit them. Such inhibitory activity—often mediated by an amino acid known as GABA—helps to shape the neural networks that underlie all behavior

by David I. Gottlieb

One usually conceives of the activity of the nervous system as a pattern of excitations. Sensory stimuli are converted into impulses that are relayed from nerve cell to nerve cell before ultimately being converted into responses. By and large this picture is an appropriate one. Yet there is an entire class of operations in the nervous system that are not excitations at all but inhibitions, and these operations are mediated by a special class of inhibitory neurons (nerve cells). Rather than exciting its target cell, the firing of an inhibitory neuron damps the target's firing or eliminates it altogether. The constraining action is accomplished by the release of specific molecules called inhibitory transmitters. Among the most prevalent of them is a simple amino acid called gamma-aminobutyric acid, or GABA; neurons that secrete GABA are referred to by neurobiologists as GABAergic.

GABA was established as an inhibitory neurotransmitter by a long line of experiments begun in the 1950's. Within the past five years the techniques of molecular biology have made great new strides possible. By means of those techniques much has been learned about GABA, about the enzyme that makes it and about the receptor to which it binds. At the same time the function of inhibitory neural networks is becoming clearer. Not only do such circuits act as brakes on the entire nervous system—preventing a runaway spree of neural firing—but also they help to "tune" the specific responsiveness of the excitatory networks that convey and interpret information about the external world. Indeed, as more is understood about inhibitory networks, they seem increasingly to be equal partners with the excitatory circuitry that generally dominates one's thinking about the activity of the nervous system.

In the investigation of the nervous system most of the attention was initially concentrated on the excitatory pathways. By the 1940's a considerable amount was known about how the excitatory impulse is generated and conveyed from neuron to neuron. An excitatory signal travels along a neuron as a minute electrical change. Ordinarily the inside of a neuron has a net negative electric potential with respect to the outside of the cell; the difference is maintained by pumps and channels that distribute electrically charged ions (such as those of sodium, potassium and chlorine) differently on the inside and outside of the cell. As the electrical impulse, or "action potential," passes along the neuron, channels open and close, allowing ions to flow, and the inside of the cell briefly becomes positive with respect to the outside before returning to the resting level.

The action potential travels away from the body of the cell along an axon (the type of fiber that sends nerve signals). At the end of the axon it reaches the synaptic terminal, a bulbous structure forming the "near" side of the synapse, or gap, between nerve cells. The arrival of the action potential causes a rapid discharge of neurotransmitter from the terminal into the synaptic cleft. The transmitter diffuses across the synapse and binds to receptor molecules on a dendrite (receiving fiber) of the second neuron. The interaction of transmitter and receptor elicits a new electrical signal in the dendrite. Synapses are so well adapted to their function that this complex process takes place in about a thousandth of a second.

By the 1950's, however, it had become clear that not all synapses work this way: many of them block the activity of the postsynaptic neuron. It was reasonable to assume that specialized transmitters are responsible for the inhibitory effect. Quickly evidence began to accumulate suggesting that GABA is one such molecule. In the early 1950's Eugene Roberts of Washington University in St. Louis, Jorge Awapara of the University of Texas and Sidney Udenfriend of the National Heart Institute independently discovered that GABA is present in high concentrations in mammalian brain tissue. The amino acid was undetectable in other organs, which implied that it has a specific role in the central nervous system. The complexity of the mammalian brain, however, made it difficult to specify what that role might be.

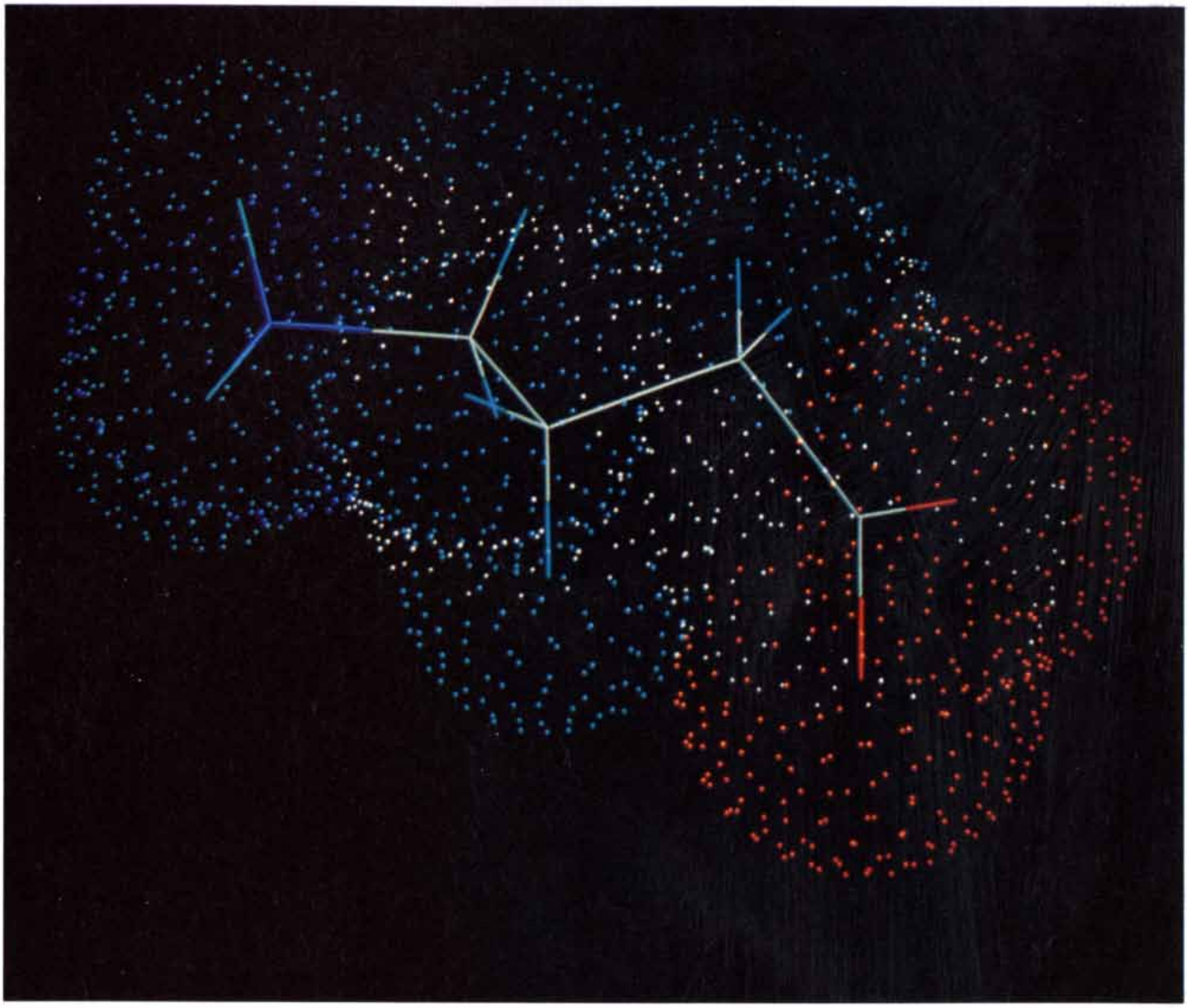
Fortunately a workable model was soon provided by simpler organisms: the crustaceans. The muscle fibers of lobsters and crayfish receive three different neural inputs. Excitatory axons extend from the central ganglia (the nodes of neural tissue along the midline of the body that serve as switching points for the nervous system) to the muscle fibers; activation of the excitatory neurons causes the muscle to contract. A second type of neuron runs from the muscle fiber to the central ganglia. Called the stretch-receptor neuron, it conveys information about muscle length to the central nodes. Finally, there are inhibitory axons that are connected to both the muscle fibers and the stretch-receptor cells. When the inhibitory neurons fire, they suppress the activity of the muscle and the stretch receptors.

Several lines of evidence demonstrate conclusively that GABA is the inhibitory neurotransmitter of this system. Ernst Florey of McGill University showed that GABA, applied to a preparation of isolated muscle fiber and neurons, suppresses the discharge of the stretch receptor; others showed that GABA does the same for muscle fibers. A plant alkaloid called picrotoxin, which blocks the action of GABA, also blocks the action of the inhibitory axons specifically and reversibly. What is more, both GABA and the enzyme that makes it—glutamic acid decarboxylase (GAD)—are found in the inhibitory axons but not in the excitatory ones. The distribution of the enzyme was uncovered by a team from the Harvard Medical School that was led by Stephen W. Kuffler and Edward A. Kramt and also included J. Dudel, David D. Potter and Zach Hall.

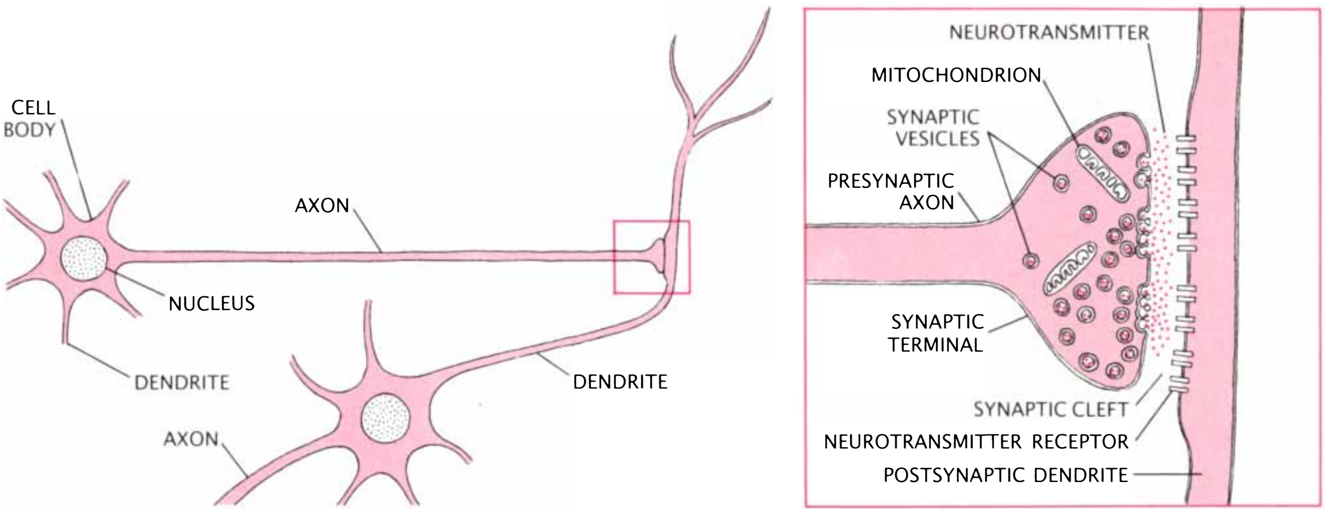
The crowning touch was then provided by the Harvard investigators, who dissected the crustacean neuromuscular preparation free of surrounding tissue and placed it in a saline bath. Stimulation of the inhibitory axons, but not the excitatory ones, caused the release of GABA into the bath. Taken together, these results clearly showed that GABA is the inhibitory transmitter in the crustacean system. The wide-ranging series of experiments carried out by the Harvard workers served as an often emulated model for establishing the nature of the transmitter at other synapses, and by the 1960's the focus of the attack had shifted back to the mammalian brain.

A wealth of evidence from many areas of the brain confirms that GABA is an inhibitory transmitter in the brains of all mammals, human beings included. One of the areas where GABAergic neurons have been best studied is the cortex (outer layer) of the cerebellum. The cerebellar cortex is responsible for the smooth coordination of muscular activity, a job it carries out by influencing neurons in various higher-brain structures. Of the neurons in the cerebellar cortex only the one called the Purkinje cell has an axon that leaves that structure: the Purkinje cell axons terminate in structures called the deep cerebellar nuclei (DCN) that lie underneath the cortex. As a result of this arrangement the cortex can influence other brain structures only through the action of Purkinje cells.

In the early 1960's Masao Ito of



GAMMA-AMINOBUTYRIC ACID (GABA) is an amino acid that serves as a neurotransmitter, or carrier of signals between neurons (nerve cells). The computer image, made by Garland R. Marshall of the Washington University School of Medicine, shows the atomic structure of GABA (*lines*) and the distribution of the surrounding electric charge (*clouds of dots*). Yellow-white corresponds to carbon, blue-green to hydrogen, blue to nitrogen and red to oxygen. GABA is an inhibitory neurotransmitter: when it is released into a synapse (the gap between nerve cells) by one neuron, it prevents the neuron on the other side from firing.



NERVE SIGNAL IS TRANSMITTED across the synapse that connects two neurons. The signal, a minute electrical disturbance known as the "action potential," travels away from the cell body down a long fiber called the axon (*left*). At the end of the axon it reaches a bulbous structure called the presynaptic terminal. Vesicles in the terminal release a neurotransmitter that diffuses across the synaptic cleft and binds to receptors on the postsynaptic dendrite, a receptive fiber, eliciting a new electrical signal.

the University of Tokyo and his colleagues made a surprising discovery. They found that stimulating the Purkinje cells caused the rate of firing of the neurons in the DCN to decrease rather than increase. Their conclusion was that the Purkinje cells' action is inhibitory. Since the Purkinje cells provide the sole route from the cerebellar cortex to the underlying nuclei, it follows that the cortex exercises all its effect on the rest of the brain through inhibition.

It was not long before it was found that the transmitter that mediates the inhibitory effect is GABA. The evidence followed the lines that had been established in the crustacean work. When GABA is applied to the cells of the DCN, their firing is blocked. Agents such as picrotoxin and another substance called bicuculline, which block the effect of GABA, also block activity in the synapses made by Purkinje cells. The Purkinje cells contain high levels of GAD and GABA. Demonstrating that GABA is released when Purkinje cells fire was the most difficult task, but K. Obata and K. Takeda of Tokyo Medical University found an elegant way to measure the release of transmitter into the fourth ventricle (the large, fluid-filled cavity above which the DCN lie). They showed that the firing of Purkinje cells is indeed accompanied by the release of GABA.

Once it had been clearly demonstrated that GABA is an inhibitory transmitter in the mammalian brain, work progressed rapidly, yielding a picture of the "typical" GABAergic neuron. GABAergic neurons contain high concentrations of the synthesizing enzyme GAD. The enzyme is found throughout the cell, but it is particularly concentrated in the presynaptic terminal. Within the terminal are many vesicles, and it is thought that the stored GABA is released from the vesicles, although firm evidence is still lacking on this point. In addition, the outer membrane of the terminal incorporates molecular pumps that help to clar the synapse of GABA, thereby preparing the synapse for the next firing.

To have its effect, GABA not only must be released but also must be bound on the postsynaptic side of the synapse by receptors: molecules specific for that purpose embedded in the outer membrane of the postsynaptic cell. In the case of GABA there are two such receptors, designated $GABA_A$ and $GABA_B$. Each receptor binds GABA and then produces a change in ion permeability. In the case of the *A* receptor the permeability of the membrane to chloride ion is increased; in the case of the *B* receptor it is the potassium permeability that increases. In both instances the effect is the same: the potential difference between the inside and the outside of the postsynaptic cell increases, and so the cell becomes less likely to fire.

GABAergic neurons are widespread in the central nervous system. Almost every major division of the brain and spinal cord includes some of them. In many regions they make up an appreciable minority of all neurons: between 20 and 40 percent. Remarkably, in some regions these inhibitory neurons actually make up the majority of all the nerve cells that are present.

On the basis of their anatomical form GABAergic neurons have been divided into three groups. Type I neurons have many processes, or extensions, but the processes are not clearly divided into dendrites and axons, as is the case for most nerve cells. Instead each process is both a sender and a receiver of messages. An example is provided by the granule cells of the olfactory bulb, which outnumber all other neurons there and inhibit specific neighboring cells.

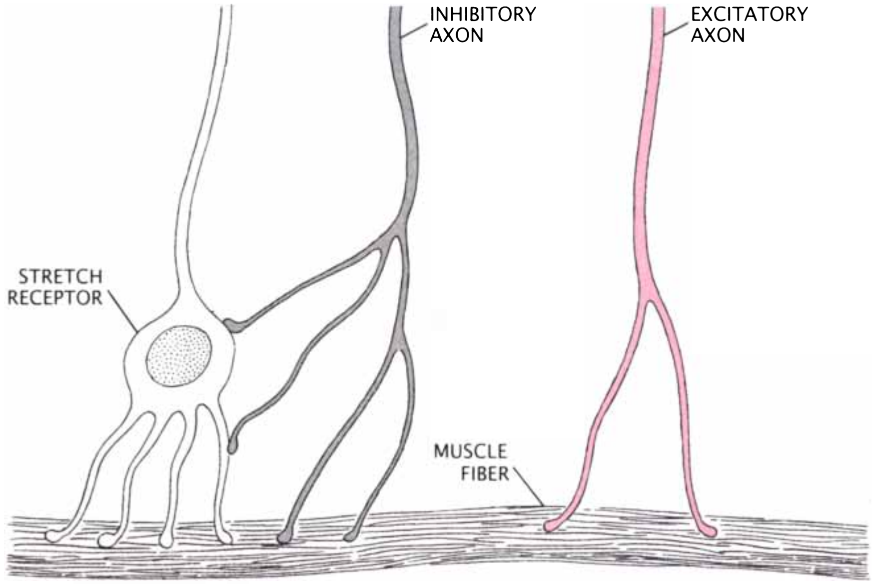
In Type II and Type III neurons, on the other hand, dendrites and axons are clearly distinguished. All Type II cells send messages to other neurons within the adjacent gray matter of the brain (which consists of nerve bodies, nerve fibers and supportive tissue), but they can vary widely in the number of cells with which they make contact. For example, the basket cell (a Type II neuron in the cerebellum) generally synapses with about six Purkinje cells; its neighbor, the Golgi cell, however, may make connections with up to 10,000 of its target cells. Type III neurons are called projection neurons because their axons leave the gray matter and enter the white matter (which consists largely of nerve-fiber bundles), thereby serving to inhibit neurons in distant brain structures.

It is not yet clear how the anatomical form of each GABAergic neuron is correlated with its function, and the problem is under active investigation. In the meantime a considerable amount has been learned at the molecular level about how GABAergic synapses operate. Such knowledge has been gleaned partly at the intersection of pharmacology and molecular biology. As it happens, the pharmacological insights were gained first. The 1960's was the decade when the benzodiazepines came into widespread use. This class of chemicals, whose best-known member is marketed under the name Valium, is prescribed to relieve anxiety, pain and muscle spasms, to induce sleep and for acute (but not chronic) control of epilepsy.

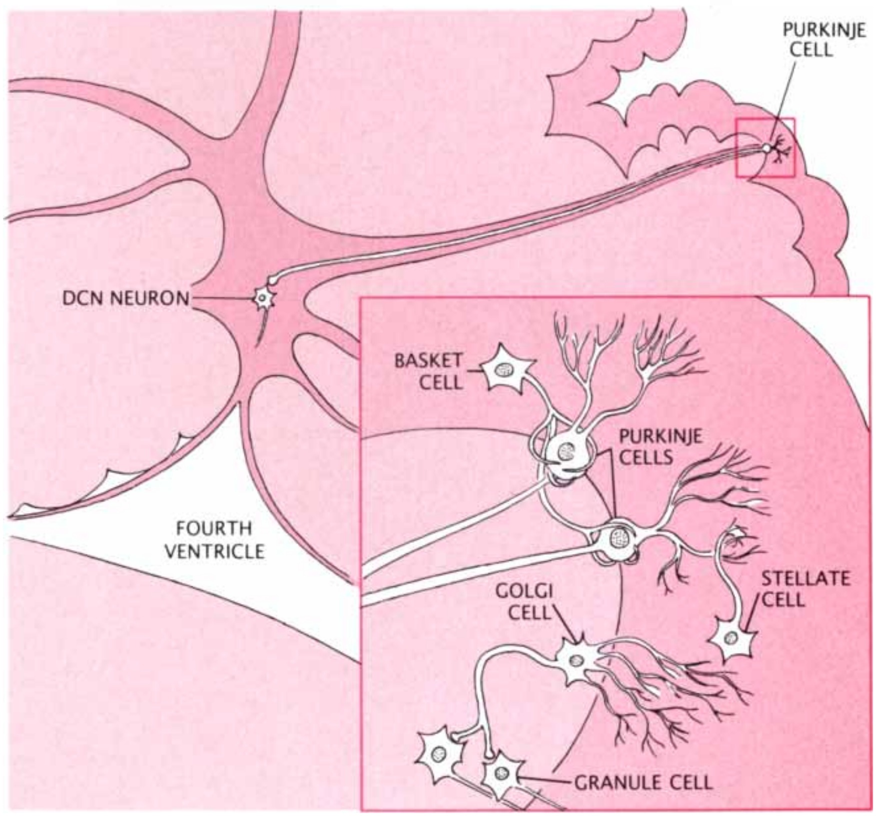
Quite early, Erminio Costa of the National Institute of Mental Health showed that the benzodiazepines can have such varied effects because they potentiate (augment) the action of GABA. Further work by Gerald D. Fischbach, Dennis W. Choi and David H. Farb of the Harvard Medical School and Jeffery L. Barker and Robert Macdonald of the National Institute of Mental Health demonstrated that the augmentation takes place at the level of the GABA receptor. As I have described above, one type of GABA receptor controls chloride permeability. The Harvard and NIMH groups showed that the benzodiazepines lower the concentration of GABA needed to increase chloride permeability. Significantly, the drugs do not affect permeability when they are administered alone: they can only intensify the effect of the normal transmitter. It is now known that the benzodiazepines bind to the GABA receptor at a site different from the site where GABA binds, but their precise mechanism is not understood.

In the 1980's, studies of GABAergic neurons have expanded to include molecular biology along with pharmacology. Even in molecular-biological studies, however, the benzodiazepines provided a crucial foothold, as I shall describe.

Until recently students of the central nervous system felt as if they were to some extent excluded from the unfolding drama of molecular biology. The great complexity of the mammalian brain made neurobiologists despair of learning anything by the rapidly advancing methods of molecular genetics. In the past few years, however, it has become possible to analyze key brain genes and proteins with the full power of molecular biology. That power has been applied to both GAD, the enzyme that



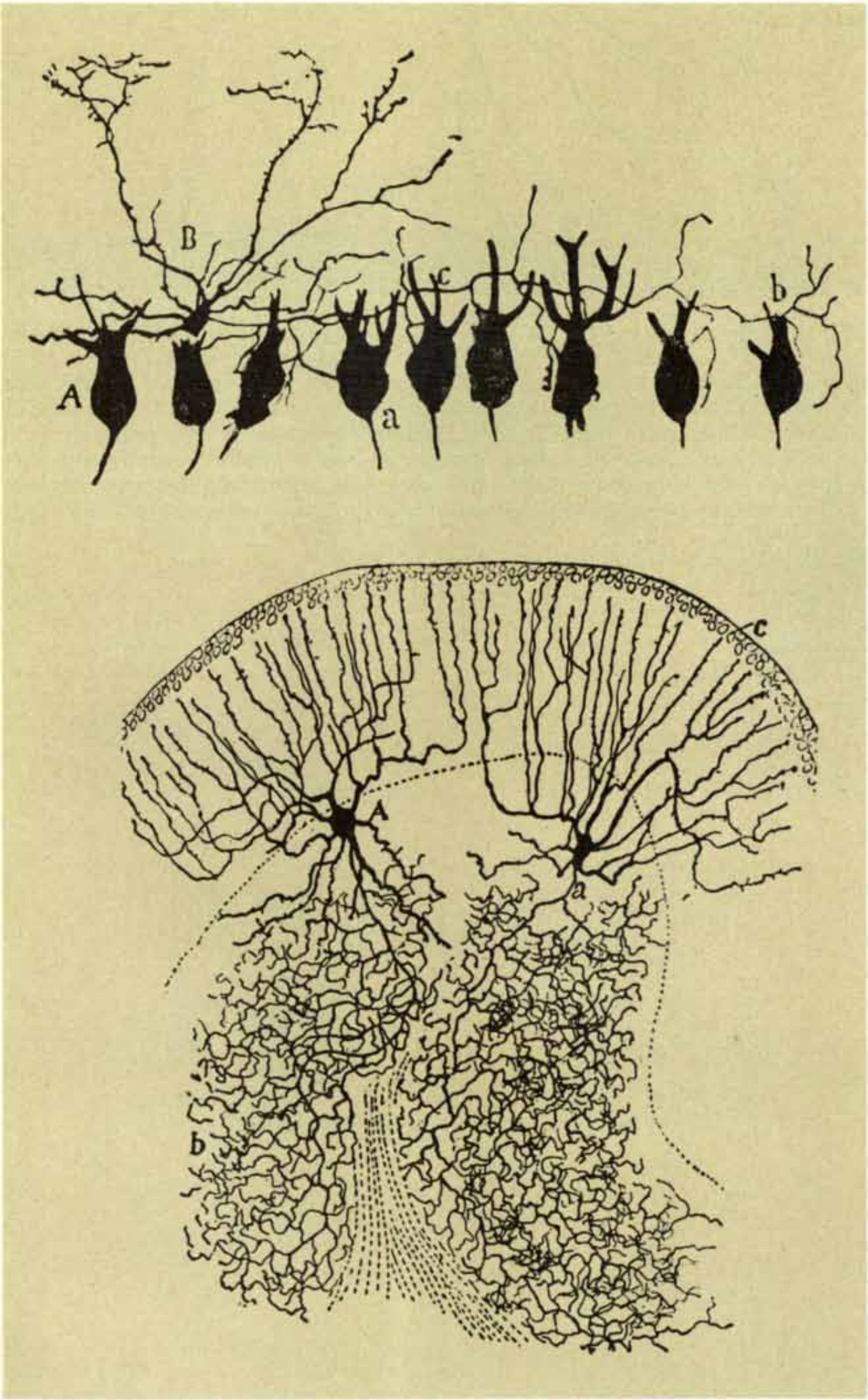
SIMPLE EXPERIMENTAL SYSTEM consisting of nerve and muscle fibers from a crustacean (such as a crayfish or a lobster) provided the basis for proving that GABA is an inhibitory transmitter. The excitatory axon, coming from the main nerve bundles near the midline of the body, causes the muscle fiber to contract. The stretch receptor conveys information about the contractional state of the muscle. The inhibitory axon blocks muscular contraction and the firing of the stretch receptor. In the 1960's several lines of evidence (including the fact that blockers of GABA readily and reversibly cancel the inhibitory effect) showed that GABA is the inhibitory transmitter in this system.



CORTEX OF CEREBELLUM exercises its effects on other brain structures solely by means of inhibitory neurons. The cortex, shown in a schematized cross section, is responsible for ensuring smooth coordination of the muscles. The output of the cerebellum is carried by neurons coming from structures called the deep cerebellar nuclei (DCN). All the axons extending from the cerebellar cortex to the DCN are those of Purkinje cells: inhibitory neurons whose transmitter is GABA. The Purkinje cells are part of a complex network of other neurons with both excitatory and inhibitory inputs (*inset*).

makes GABA, and the GABA receptor, with intriguing results.

The new techniques have made it possible to achieve results much more quickly than they could possibly have been achieved with the older methods of classical protein chemistry. The older methods were beset with difficulty, mainly because both GAD and the GABA receptor are present in very small quantities in the brain and are difficult to separate from the multitude of other brain proteins. In their pioneering studies on the structure of GAD, Roberts and Jan-Yen Wu of the City of Hope Medical Center were forced to liquefy more than 9,000 mouse brains to obtain enough of the enzyme to perform analytical studies. Clearly, under such conditions the work on GAD would have progressed slowly.



INHIBITORY NEURONS have a broad range of anatomical forms, as is suggested by a pair of drawings by the 19th-century Spanish histologist Santiago Ramón y Cajal. Both panels depict cells from the cerebellum. The upper one shows a basket cell (*B*) with a row of Purkinje cells (*A*). The lower one shows a pair of Golgi cells; the tangled masses below the cell bodies are the axons, which can make contact with 10,000 other cells.

One of the most effective tools in the kit of molecular biology is the monoclonal antibody: an antibody that binds specifically and exclusively to a single protein. With such an antibody in hand one can separate a protein from its environment, no matter how complex that environment is. Recently my colleagues Yen-Chung Chang and James Schwob and I succeeded in preparing a monoclonal antibody to GAD. Having such a monoclonal antibody makes it possible to purify the enzyme rapidly and in high yield, which has enabled us to work out part of the amino acid sequence of the enzyme and further analyze its molecular composition.

An even more detailed look at GAD has been obtained by means of another of the ingenious tools of the new molecular biology: the complementary DNA (cDNA) clone. Such a DNA clone is called complementary because it matches the messenger RNA (mRNA) for a particular protein, in this case the enzyme GAD. Because the cDNA carries the genetic information that specifies the string of amino acids in the protein, analysis of the clone yields the complete sequence of amino acids, a most valuable piece of information. Obtaining the clone, however, is no small feat. In this case the job was done by Daniel L. Kaufman and Allan J. Tobin of the University of California at Los Angeles.

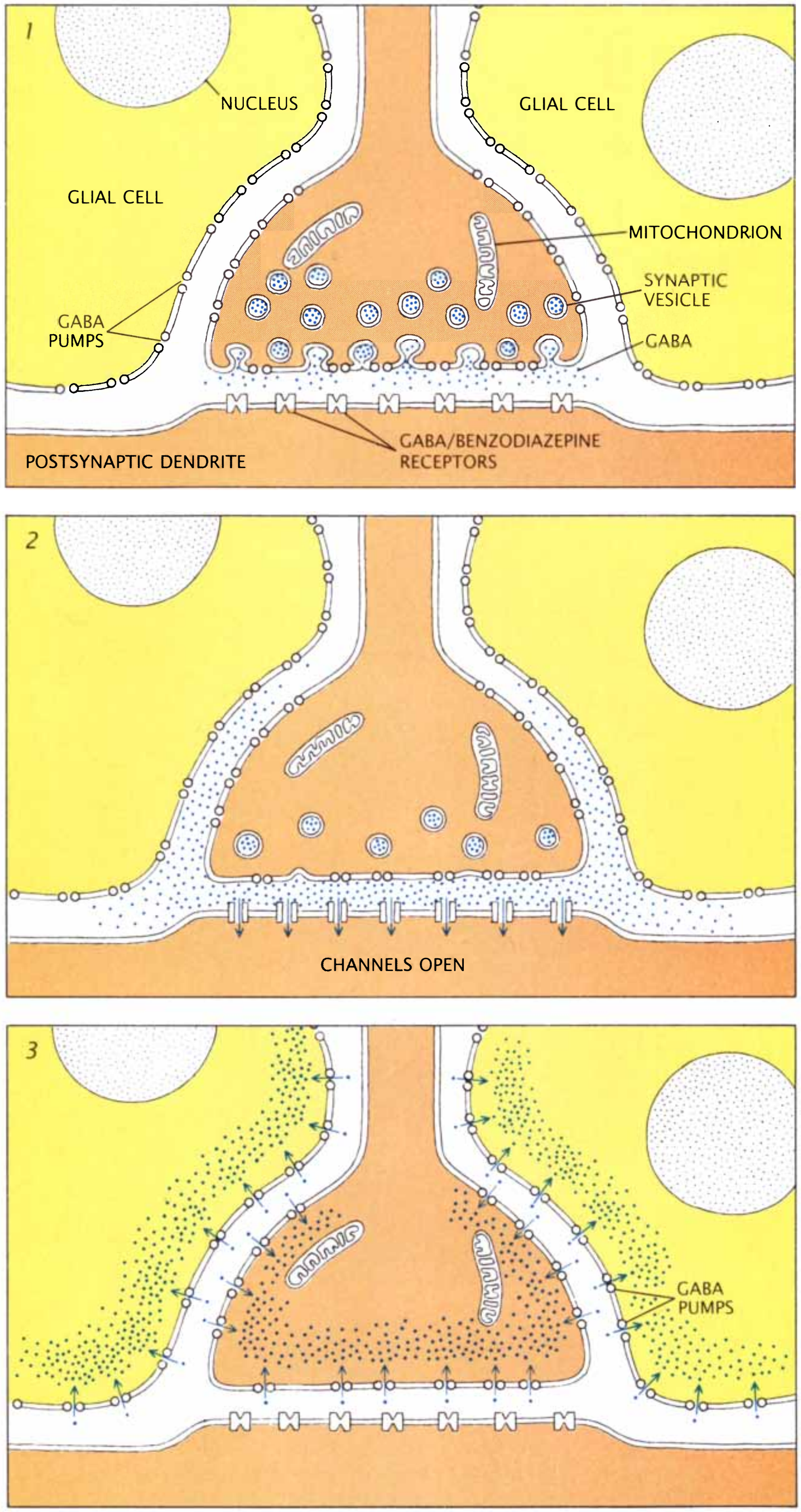
Kaufman and Tobin first prepared a cDNA "library" from the brain of a cat. The library includes cDNA's that are complementary to all species of messenger RNA found in the cat's brain. The next step was to select, from among all the DNA's in the library, the one corresponding to the desired enzyme. In order to do so, the library was inserted into a simple virus called a phage that infects bacteria; one small piece of the library was put into each phage. Once inside the bacterium, the DNA carried by the phage is expressed, or used to make protein. When that happens, antibodies to GAD can be used to pick out the bacteria carrying the DNA for the enzyme. Tobin's group sequenced the clone and deduced the entire amino acid sequence of GAD.

The GABA receptor is an even more elusive protein than the enzyme that makes GABA: it accounts for only one part in 50,000 of all brain protein. Yet a series of experiments that entailed a crucial exploitation of the benzodiazepines has made it possible to understand the basic structure of the receptor.

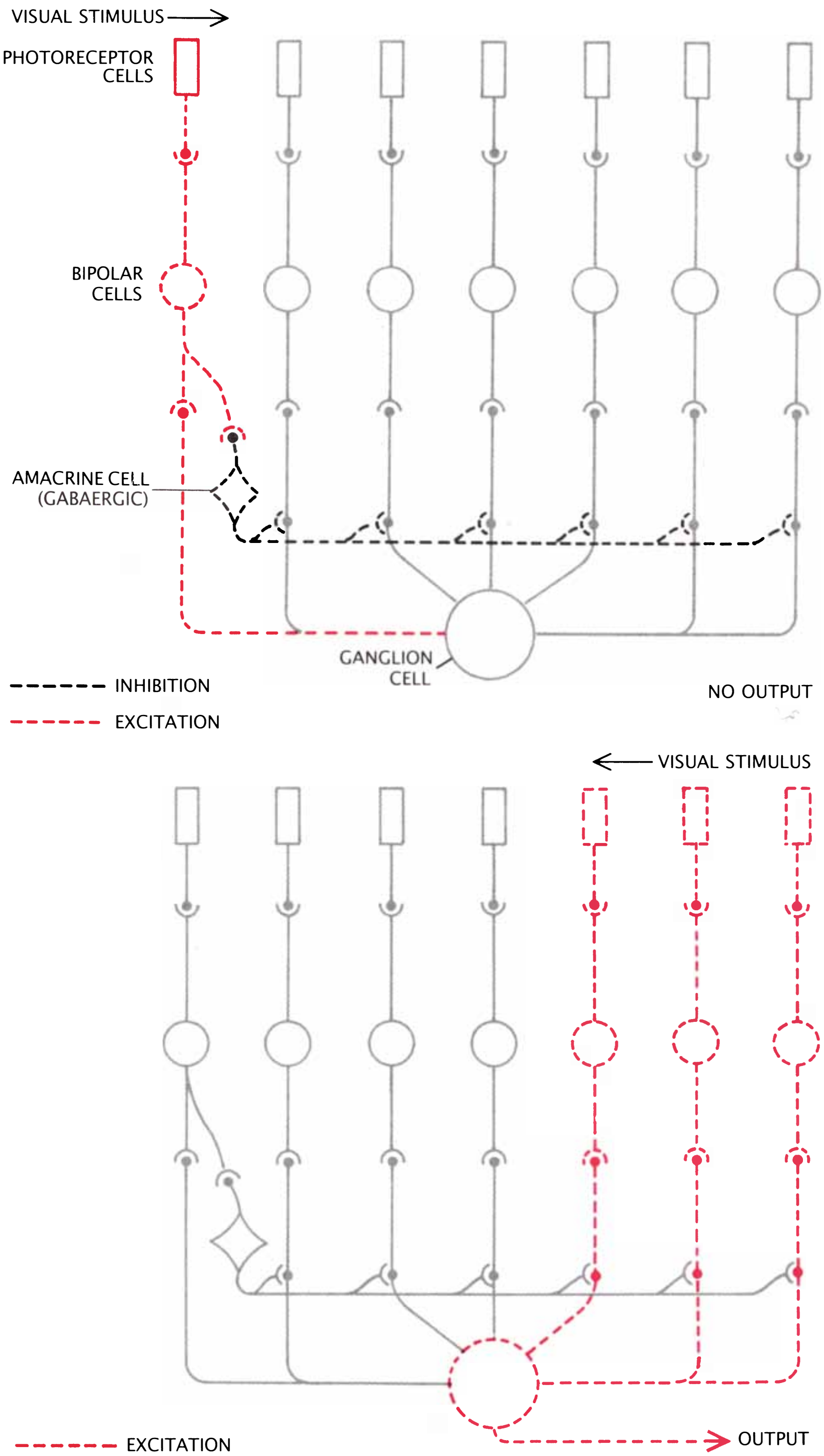
The GABA receptor is an integral membrane protein: it extends through the bilayered outer membrane of the postsynaptic neuron. Richard W. Olsen of the University of California at Los Angeles first showed that the intact and functional receptor could be separated from the membrane by means of detergents. Building on Olsen's work, Eric A. Barnard of the U.K. Medical Research Council's Molecular Neurobiology Unit and Hans Möhler of the Roche Company in Switzerland have been able to separate the receptor from the mass of other proteins embedded in the membrane. After being released from the membrane with detergent, the proteins were passed through a column packed with minute beads to which benzodiazepine molecules were bound. The benzodiazepines hooked the passing GABA receptors, which were then stripped from the column so that they could be analyzed chemically.

The benzodiazepines, it should be noted, bind only to the $GABA_A$ receptor. Hence the receptor molecule removed from the column was the *A*-type receptor, which, as I mentioned, affects the permeability of the membrane to chloride ions. The analysis carried out by the British and Swiss groups showed that the $GABA_A$ receptor consists of two conjoined subunits, one of about 55,000 daltons and the other of about 50,000 daltons. (A dalton is a measure of molecular weight equal to the weight of a single atom of hydrogen.) What is more, having the purified receptor molecule made it possible to prepare monoclonal antibodies and take other steps that laid the groundwork for cloning the receptor gene.

Just last July, Barnard's group reported that they had cloned the genes for the $GABA_A$ receptor. Several significant results followed from their breakthrough. First, as in the case of GAD, the genes yielded the complete amino acid sequence of the paired receptor subunits. Knowledge of the sequence provided the basis for formulating a model of the receptor's geometry and position in



SYNAPSE MUST BE CLEARED to convey the next signal, a job done by pumps in the membrane of the axon and the surrounding glial cells. On being released by vesicles (*1*), GABA crosses the synapse and binds to specific receptor molecules in the membrane of the postsynaptic dendrite (*2*). The binding changes the membrane's permeability to such ions as chloride and potassium. (The ion flow is what generates the initial electrical impulse.) Molecular pumps then quickly begin removing GABA from the synapse (*3*).



NEURAL NETWORK can be adjusted in its sensitivity by a GABAergic neuron. The illustration shows schematically a network that responds to visual stimuli and forwards information to the brain. Photoreceptor cells in the retina respond to changes in light intensity and pass their signal to bipolar cells. Inputs from many bipolar cells are gathered and relayed by ganglion cells. If a stimulus moves from left to right (*top*), the first photoreceptor cell excites a bipolar cell, which triggers a GABAergic amacrine cell. The GABAergic neuron silences the other bipolar cells. The bipolar cell also excites the ganglion cell. Many such excitatory inputs, however, are required for the ganglion cell to fire; the single input is not sufficient. Hence there is no ouput. The bipolar cells on the other side of this network, however, are not wired to the amacrine cell. Therefore a stimulus that is moving from right to left (*bottom*) does yield an output to the brain.

the membrane. The structure reported by Barnard's group in their paper includes a set of eight helical regions spanning the membrane (four per subunit) connected by linear stretches. The helical units probably form the channel through which the chloride ions pass.

Not content with that substantial advance, Barnard and his co-workers took two further steps. The first was to inject mRNA's corresponding to the receptor genes into oocytes (eggs) from a frog of the genus *Xenopus*. In the egg the RNA's made their way to ribosomes, where they were translated into receptor molecules. That the receptors were intact and functional was shown by the fact that they responded normally to GABA and to the benzodiazepines when these substances were applied to the surface of the oocytes. This remarkable demonstration raises considerably the reliability of the proof that the correct stretches of DNA had indeed been cloned.

The second step was to show that there are strong homologies (similarities in DNA sequence) between the receptor for GABA and the receptor for acetylcholine, which is an important excitatory transmitter. The conclusion drawn by Barnard and his colleagues is that both receptors belong to a "superfamily" of receptor molecules, which is very old in evolutionary terms. One intriguing implication of this hypothesis is that the receptors for both excitatory and inhibitory transmitters may well be descended from a common molecular ancestor.

Thinking about the evolutionary descent of inhibitory neurotransmitters naturally raises the question of their function. Clearly, inhibitory activity, which is so widespread in the brain and spinal cord, must fulfill a significant function. It has long been thought that the inhibitory neurons act collectively as a governor that prevents excitatory neurons from firing to excess. There is some evidence that the inhibitory nerve cells do indeed carry out this function. For example, the administration to experimental animals of agents that block the effect of GABA (such as picrotoxin or bicuculline) triggers widespread excess neural activity accompanied by convulsions.

Yet this rather passive, nonspecific activity is certainly not the whole story, as some interesting recent work has shown. Since picrotoxin and bicuculline selectively block GABAergic synapses, it is possible to apply these compounds and examine how neural networks operate with their GABAergic synapses silenced. My colleagues John H. Caldwell, Harry J. Wyatt and Nigel W. Daw took that approach with the rabbit retina. It had already been shown by Horace B. Barlow of the University of California at Berkeley that many of the ganglion cells in the rabbit retina—which are the cells that relay visual information to the brain—are directionally selective: they respond to a stimulus moving, say, from left to right but not to one moving from right to left.

Daw and his co-workers examined the responses of such ganglion cells before, during and after the application of picrotoxin. They found that in the presence of picrotoxin the directional sensitivity was erased—the ganglion cells responded equally well to a stimulus moving in either direction—only to reappear when the drug was washed out of the tissue.

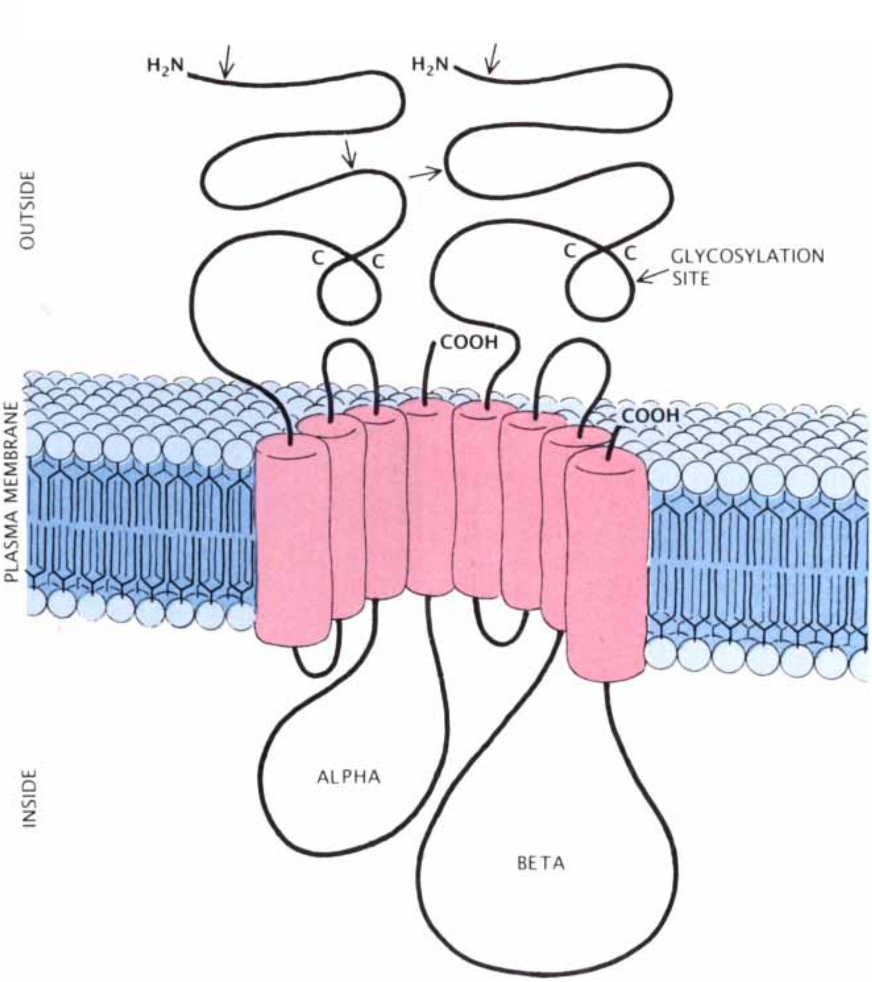
The most plausible explanation of these remarkable results is that a GABAergic neuron endows the ganglion cell with its directional sensitivity. Light striking the retina activates photoreceptors (called rods and cones) that trigger the ganglion cells only through the intermediate effects of cells called bipolar cells. Branches of the bipolar cells also excite GABAergic neurons called amacrine cells, whose processes are arranged horizontally and make contact with the ganglion cells. A certain subset of amacrine cells must be arranged asymmetrically, so that when they are stimulated, they inhibit ganglion cells to the left or the right, but not both [*see illustration on opposite page*]. In the presence of picrotoxin, the amacrine cell's effect is blocked, and the ganglion cell responds to movement in either direction.

Thus the amacrine cell gives the rabbit's retina a key feature: the capacity to detect the direction of movement. It is not difficult to see how such sensitivity could be of great evolutionary benefit in judging the motions of predators or other objects. Similar benefits may come from another effect of the GABAergic neurons, which has been examined by Robert W. Dykes of McGill University and his colleagues, who worked with cats.

Healthy cats are capable of distinguishing very well between stimuli quite close together on their skin. They are able to do so because certain neurons in the brain respond to stimuli only within a highly circumscribed area, which is called their receptive field. These neurons are concentrated within a structure called the somatosensory cortex. When bicuculline is injected into the somatosensory cortex, the receptive fields of many of the neurons expand dramatically beyond their normal limits. As in the case of the rabbit retina, the effect is reversible, disappearing after the drug is removed.

The best explanation for these effects is, again, that GABAergic neurons are selectively canceling sensory inputs. It would seem that neurons in the somatosensory cortex actually receive excitatory inputs from a large area on the skin surface. Not all the inputs from that area, however, are created equal. Those from the peripheral region are accompanied by inhibitory inputs that cancel their effects, whereas those from the central region pass through uninhibited, thereby defining the small region of sensitivity.

These two examples show clearly that the role of inhibitory neurons in general, and GABAergic neurons in particular, goes far beyond regulating overall levels of activity in the manner of a neuronal governor. In their newly understood capacity as specifiers of responsiveness they may be of fundamental significance to the operation of the brain. Modern concepts of brain function stress that intricate neural networks are the basis of the brain's capacity to sift and analyze the information supplied by the senses. By tuning the responsiveness of key neurons, the GABAergic neurons may shape the networks, and so, as much as the excitatory pathways, form the basis of sentience and action.



GABA RECEPTOR twines through the outer membrane of the postsynaptic cell. The receptor consists of two subunits, designated alpha and beta. Each subunit is a long chain of amino acids that includes four helical regions (depicted as cylinders) spanning the membrane and linear regions extending beyond it. The helical segments probably form a channel for chloride ions. The arrows indicate points where sugar molecules are attached to the amino acid chain. This model is for the receptor known as $GABA_A$. It was proposed in 1987 by Eric A. Barnard of the U.K. Medical Research Council's Molecular Neurobiology Unit and his co-workers, who cloned the genes for the receptor molecule.

How Climate Evolved on the Terrestrial Planets

Planets with temperate, earthlike climates were once thought to be rare in our galaxy. Mathematical models now suggest that if planets do exist outside the solar system, many of them might be habitable

by James F. Kasting, Owen B. Toon and James B. Pollack

Why is Mars too cold for life, Venus too hot and the earth just right? At first glance the answer to this question, the so-called Goldilocks problem of climatology, may seem simple. Common sense suggests that the earth, with a livable mean temperature of 15 degrees Celsius, just happened to form at the right distance from the sun, whereas Mars (–60 degrees C.) and Venus (460 degrees C.) did not; as a result only on the earth does one find the liquid surface water that is crucial for life.

Yet happenstance is not the complete explanation for the temperatures of these terrestrial, or rocky, planets. We propose that the three neighbors, all of which formed when large numbers of bodies known as planetesimals collided, were once alike in many ways. They had similar minerals on their surfaces and similar gases in their atmospheres (including carbon dioxide and water vapor), and they were all temperate enough to maintain liquid water on much of the surface. They acquired dramatically different climates largely because they differed in their ability to cycle carbon dioxide between the crust and the atmosphere. Carbon dioxide, like water vapor and certain other substances, is a "greenhouse" gas: it allows sunlight to pass through it, but it absorbs infrared radiation (heat) that rises from the planet and reradiates part of this heat back to the surface.

More specifically, calculations by our group at the National Aeronautics and Space Administration and by other investigators suggest that the earth has always had a moderate climate primarily because its cycling mechanism increases the amount of carbon dioxide in the atmosphere when the surface of the planet cools and reduces the amount when the ground temperature rises. Mars is now frozen because it has lost the ability to cycle the gas back into its atmosphere, and Venus is a hothouse because it developed the opposite problem: it has no way of removing carbon dioxide from its atmosphere. (Mercury, the other terrestrial planet, has no atmosphere; its temperature is controlled exclusively by the sun.)

The Faint-Young-Sun Paradox

Our interest in the role of carbon dioxide in the evolution of the earth, Mars and Venus had its roots in another cosmological puzzle relating to the origin of the earth: the faint-

VENUS, EARTH AND MARS (*left to right*), shown roughly to scale, may each have been temperate enough in their early history to support life-giving, fluid water on their sur-young-sun paradox. Virtually every model of stellar evolution indicates that the sun was between 25 and 30 percent dimmer when the solar system formed some 4.6 billion years ago than it is today. Since then the solar luminosity, or intensity, has apparently increased approximately linearly with time.

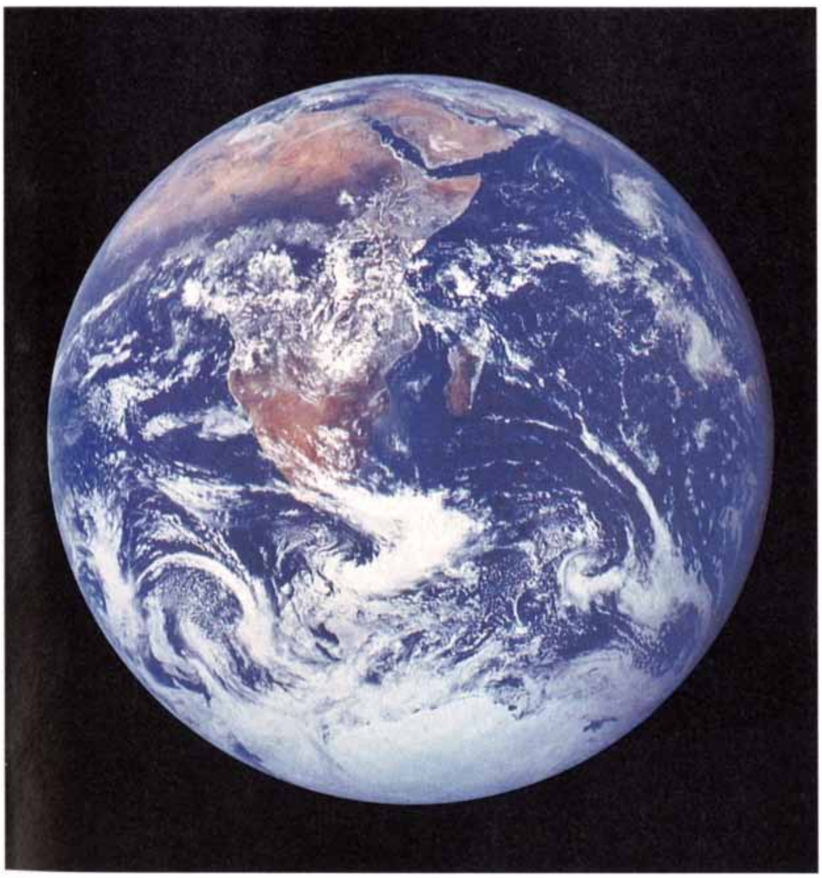
The paradox arises, as Carl Sagan and George H. Mullen of Cornell University pointed out about 15 years ago, when one realizes that if the earth's early atmosphere was the same as it is now, a weak sun would have resulted in an ice-covered earth until about two billion years ago. Yet the planet did not freeze. In fact, evidence from sedimentary rocks indicates that the earth has had liquid oceans since at least 3.8 billion years ago, when the geological record begins. Moreover, life has been present for at least the past 3.5 billion years, demonstrating that the earth's surface has never been entirely frozen during that time. (Water can remain fluid as long as the temperature is between zero and 374 degrees C.; it boils and evaporates at 100 degrees C. at sea level today but will stay liquid at higher temperatures if the atmospheric pressure is increased.)

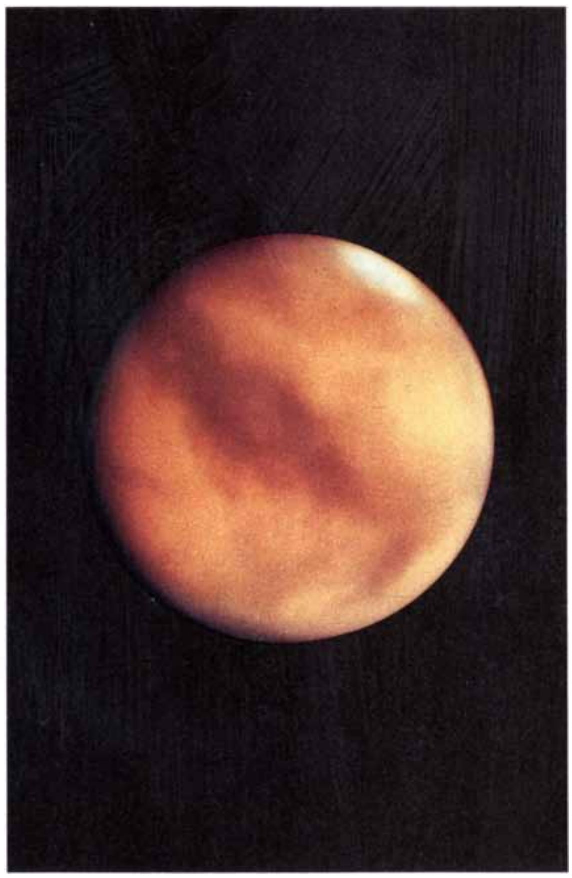
Sagan and Mullen realized that the paradox disappears if one assumes the earth's atmosphere has changed in the course of time. For instance, if the young planet had fewer clouds than it has today, less of the sunlight that impinged on the earth would have been reflected back into space, and the planet would have been correspondingly warmer. Some 30 percent of the sunlight that currently reaches the top of the atmosphere is returned to space, most of it by clouds. A chillier earth might well have had fewer clouds but the geological record suggests the early earth was actually warmer than today's. Parts of the planet are covered with glaciers now, but there is no evidence of similar glaciation before about 2.7 billion years ago.

A more probable explanation is that the greenhouse effect was more pronounced in the distant past. Sagan and Mullen suggested that ammonia (NH_3), an efficient absorber of infrared, could have warmed the climate if the gas represented just 100 out of every million molecules of the air. Subsequent studies have shown, however, that the sun would have rapidly converted ammonia into the nongreenhouse gases nitrogen and hydrogen unless it was continually resupplied to the atmosphere from the planet's surface.

Other investigations have focused on carbon dioxide, which sunlight does not readily decompose. Carbon dioxide is certainly abundant here; the amount now stored in the planet in carbonate rocks would exert a pressure of about 60 bars if it were released into the atmosphere. (One bar is equal to 14.5 pounds per square inch, the pressure at sea level. Today the earth's atmosphere contains about .0003 bar of carbon dioxide.) If just a few tenths of a bar of the stored carbon dioxide was originally present as a gas, its additional greenhouse warming would have compensated for the reduced sunlight.

The notion that higher carbon dioxide levels could have protected the early earth from freezing soon gave rise to a related idea: if the carbon dioxide level declined at a rate that precisely counteracted the increase in solar luminosity with time,





faces. Computer models suggest that differing abilities to cycle carbon dioxide between the atmosphere and the land—and not solely the planets' distance from the sun—helped Venus to lose its water and Mars to freeze, while the earth remained habitable.

the decline might account for the fact that the earth's temperature has always remained within reasonable limits. One investigator, Michael H. Hart of NASA, undertook to calculate such a compensatory rate.

Hart managed to work out a solution in which the levels of the gas declined approximately logarithmically with time, but his most interesting finding was that very few of his calculations succeeded. In other words, if the composition of the atmosphere had at any time changed at a rate different from his precise solution, the planet would have become unable to support life. If the carbon dioxide level had declined too slowly, the earth would have become a hothouse; if it had declined too quickly, the oceans would have frozen.

Hart did similar calculations for cases in which the distance between the earth and the sun was varied by small amounts. He found that if the earth had formed 5 percent closer to the sun, the atmosphere would have become so hot that the oceans would have evaporated, a condition known as a runaway greenhouse. Conversely, the planet would have encountered runaway glaciation if it had formed as little as 1 percent farther from the sun. Only in the relatively narrow range of orbits between .95 astronomical unit and 1.01 A.U. could one or the other of these climatic catastrophes be avoided. (One A.U. is the distance between the sun and the earth, or 149.6 million kilometers.) Hart termed this narrow band of orbital distances the continuously habitable zone (CHZ).

Hart's conclusions were unsettling because they suggested that the earth must have beaten extraordinary odds in avoiding the fate of Mars or Venus. Only within the past few years have investigators discovered the flaw in his hypothesis. A mathematical model developed by James C. G. Walker and Paul B. Hays of the University of Michigan and by one of us (Kasting) suggests that the changes in carbon dioxide concentration did not arise by sheer luck. Rather, carbon dioxide levels have probably fluctuated in response to changes in surface temperature. When the temperature goes up, atmospheric carbon dioxide levels decline, cooling the surface; when the surface cools, the abundance of atmospheric carbon dioxide increases and warms the surface. The existence of such a negative-feedback loop means that the earth probably has never been in danger of undergoing either the runaway greenhouse or the runaway glaciation postulated by Hart.

The Carbonate-Silicate Cycle

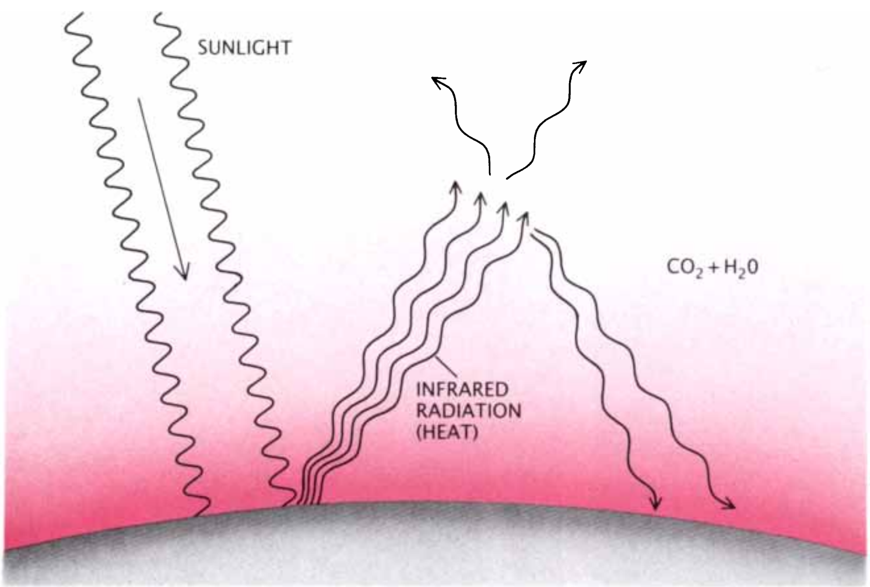
The proposed feedback system is mediated by the carbonate-silicate geochemical cycle, which accounts for about 80 percent of the carbon dioxide exchanged between the solid earth and its atmosphere over a time scale in excess of 500,000 years. The cycle begins when carbon dioxide in the atmosphere dissolves in rainwater, forming carbonic acid (H_2CO_3). The rain weathers, or erodes, rocks that contain calcium-silicate minerals (compounds of calcium, silicon and oxygen). In the process the carbonic acid reacts chemically with the rocks, releasing calcium and bicarbonate ions (Ca^{++} and HCO_3^-) into the groundwater. The water transports the ions to streams, rivers and ultimately the ocean.

In the sea, plankton and other organisms incorporate the ions into shells of calcium carbonate ($CaCO_3$). When the organisms die, they settle to the bottom of the ocean, forming carbonate sediments. As the millenniums pass, the sea floor spreads, carrying these sediments to the margins of the continents. There the sea floor slides under the landmasses and turns downward toward the interior of the planet.

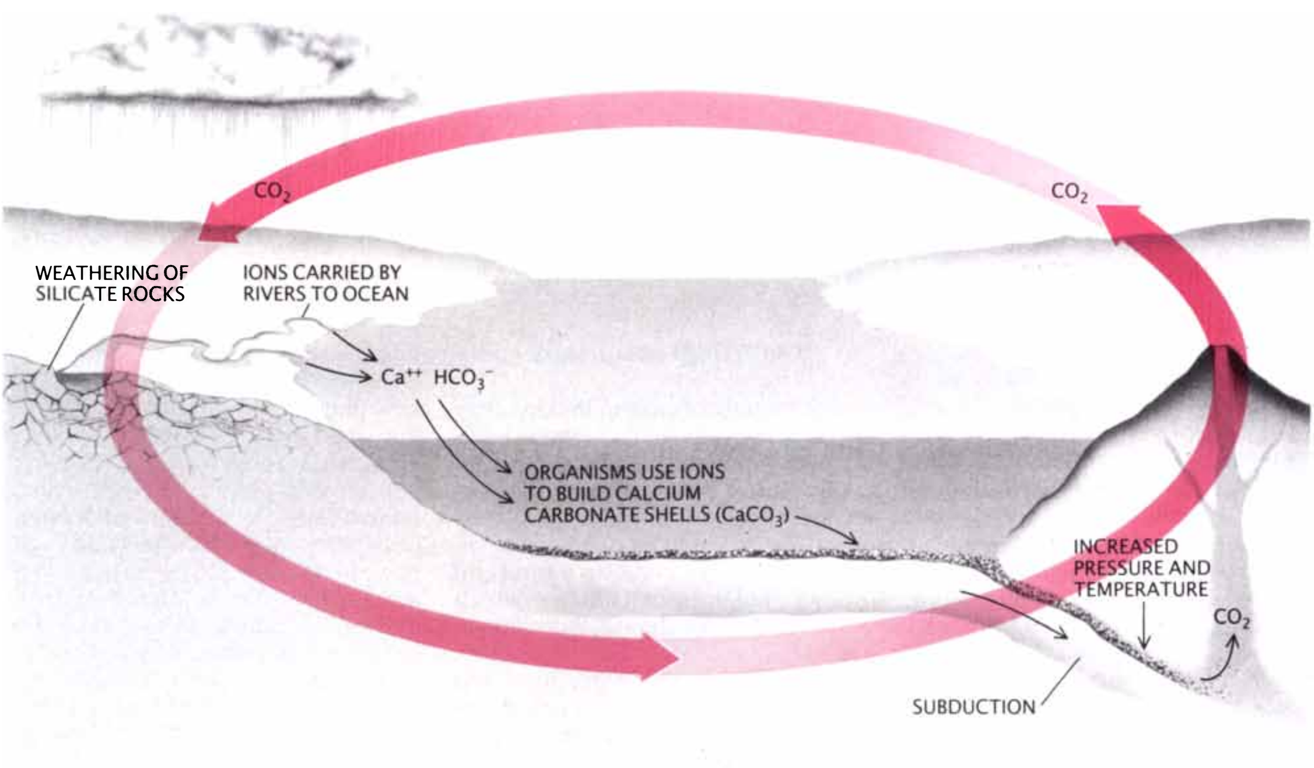
As the sediment is subducted and subjected to rising temperature and pressure, calcium carbonate reacts with silica (quartz), re-forming silicate rocks (a process known as carbonate metamorphism) and releasing gaseous carbon dioxide. The gas then reenters the atmosphere by way of midocean ridges or, more violently, through volcanic eruptions near the margins of tectonic plates.

Walker and his colleagues recognized that changes in the surface temperature in the course of time should affect the amount of carbon dioxide that leaves the environment and in turn the amount of greenhouse warming. Suppose the temperature of the surface were to fall for some reason, such as a decrease in the sun's output. When ocean temperatures fall, less water vapor evaporates into the atmosphere, and so there is less rain and therefore less weathering. Under such a circumstance the rate at which carbon dioxide leaves the atmosphere declines—but there is no change in the rate at which carbon dioxide is regenerated by carbonate metamorphism and "outgassed" into the environment. The net result is an accumulation of the gas in the atmosphere, an increase in greenhouse warming and in turn the restoration of higher surface temperatures.

Conversely, if the surface temperature were to increase for some reason, the rate of evaporation of the oceans would increase and with it the amount of rainfall. There would



GREENHOUSE EFFECT occurs when certain gases, notably carbon dioxide and water vapor, warm the surface of a planet. Such gases allow light from the sun to reach the planet, but they intercept the infrared rays (heat) that the planet radiates into space and reradiate much of this energy toward the surface. The gases raise the earth's surface temperature some 35 degrees Celsius above what it would be if they were absent.



CARBONATE-SILICATE geochemical cycle, which operates on a time scale in excess of 500,000 years, removes carbon dioxide from the atmosphere, stores it in carbonate rocks and then returns it to the air. Carbonates form when carbon dioxide dissolves in rain and reacts chemically with rocks that contain calcium-silicate minerals (compounds composed of calcium, silicon and oxygen). Such reactions release calcium and bicarbonate ions (Ca^{++} and HCO_3^-) into groundwater, which transports the ions to streams, rivers and the ocean. In the ocean, plankton and other organisms use the ions to construct shells of calcium carbonate ($CaCO_3$); they then deposit the shells in sediments on the sea floor when they die. Slowly the sea floor spreads; eventually it slips under the continents and turns downward, carrying the sediment with it. Exposed to increased temperature and pressure, the sediment gives off carbon dioxide gas, which then reenters the atmosphere primarily by way of volcanic eruptions.

be an increase in the weathering of silicate rocks and hence in the removal of carbon dioxide from the environment. The greenhouse warming would then decrease.

The feedback is perhaps easiest to visualize if one considers an extreme case. If the oceans ever froze over completely, rainfall would come to a virtual halt and carbon dioxide would build up in the atmosphere. Current rates of gas release would supply the atmosphere with a bar of carbon dioxide in 20 million years, a geologically insignificant amount of time. This amount would raise the surface temperature by about 50 degrees C.—more than enough to melt the ice and restore equable climatic conditions.

Role of the Biota

Because living organisms play an important role in the exchange of carbon dioxide with the atmosphere, some investigators have suggested that the biota are primarily responsible for modulating the climate of the earth. James E. Lovelock of the Coombe Mill Experimental Station in Cornwall and Lynn Margulis of Boston University are the main proponents of this point of view, which they call the Gaia hypothesis, after the Greek goddess of the earth. They maintain that the decrease in atmospheric carbon dioxide over the course of history has been a direct consequence of biological intervention and that without living organisms the earth's climate could well have gone the way of Mars or Venus.

The biota are certainly important. The fraction of carbon dioxide (roughly 20 percent) that does not take part in the carbonate-silicate cycle is removed from the atmosphere by photosynthetic plants. When such organisms die, they deposit organic carbon in sediments. Carbon dioxide is regenerated when tectonic processes elevate sedimentary rocks and form mountains, enabling carbon in the rocks to react with atmospheric oxygen in rainwater.

Living organisms also affect the carbonate-silicate cycle. We have discussed the role of oceanic plankton in forming carbonate sediments, but land plants may actually have a more important function. When the plants decay, oxidation of their remains enhances the abundance of carbon dioxide in the soil. As a result the concentrations of carbon dioxide in typical soils today are probably higher than they were before the appearance of vascular plants some 400 million years ago. This elevation speeds the conversion of silicate minerals into carbonate sediments.

All of this having been said, we nonetheless suggest that the fundamental controls on atmospheric carbon dioxide levels are physical rather than biological. We would argue, for example, that if the shelled organisms that now deposit calcium carbonate on the sea floor did not exist, the concentration of calcium and bicarbonate ions in seawater would rise. Once the ion concentrations reached a critical level, calcium carbonate would form without the intervention of organisms. Such must have been the case before about 600 million years ago, when the shell makers first appeared.

Similarly, calculations show that the decrease in silicate weathering caused by a complete disappearance of land plants could be offset by a temperature increase of about 10 degrees C.—a change that could be accomplished by the negative-feedback loop of the carbonate-silicate cycle. The increased greenhouse warming would produce a climate similar to that of 100 million years ago during the mid-Cretaceous period: warm, but nonetheless well suited for many forms of life, including the dinosaurs. Hence there is good reason to believe the earth would still have remained habitable even if it had never been inhabited. The carbonate-silicate cycle, acting alone, would have provided the necessary buffering mechanism.

One might well wonder whether water vapor, which today provides most of the 35 degrees C. of the earth's greenhouse warming, could have been responsible for the ability of the planet to remain temperate in the course of its evolution. The answer is no. The amount of water in the atmosphere does not counteract changes in the surface temperature. Rather it amplifies them: the abundance of water in the atmosphere increases when the surface temperature rises and decreases when the surface temperature falls. It follows, then, that only an overall decline in carbon dioxide levels can account for the fact that as the sun became brighter over the eons, the surface temperature of the earth did not rise in parallel but remained within a livable range.

Buffering Fails on Mars

The cycling of carbon dioxide may have kept the earth's climate within reasonable bounds as the planet evolved, but if a similar process ever existed on Mars, it failed to do the same for that planet. Today the atmosphere there consists entirely of just .006 bar of carbon dioxide, which provides a greenhouse warming of only about 6 degrees C.

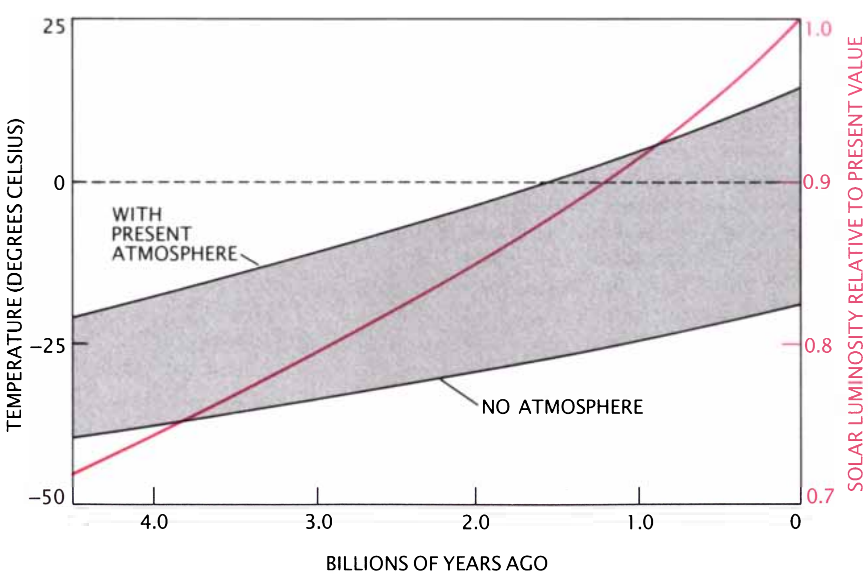
Is it possible that Mars was cold from the start and that its climate has undergone little change in the past 4.6 billion years? That is unlikely. Photographs made by NASA's Mariner and Viking spacecraft show that the Martian surface is cut by many channels that were almost certainly formed by running water. Although some of the channels could have been formed in a cold climate by the sudden release of water from great depths, the valley networks that crisscross the oldest terrain on Mars are thought to have required warmer temperatures for their genesis. The planet also had a higher rate of erosion during the first billion years of its history than it does today, according to estimates made by Peter H. Schultz of Brown University. This finding too suggests that the planet was once warm enough to support liquid water.

Geologists do not know exactly how hot Mars was, but the surface may well have been warmed by the greenhouse effect of a dense carbon dioxide atmosphere. Our calculations indicate that an atmosphere of between one and five bars of carbon dioxide would have kept parts of the Martian surface above freezing early in the planet's history. The lower figure applies to conditions at the Martian equator when the planet is closest to the sun; the higher figure is an average that applies to the entire planet.

It is within the realm of possibility that Mars once had that much carbon dioxide, even though these amounts are some 150 to 800 times more than is present in the atmosphere now. If Mars, whose mass is about a tenth of the earth's, once possessed a proportional amount of carbon dioxide, the planet would have had a total carbon dioxide endowment equivalent to about 10 bars. (To arrive at this figure one must take into account the smaller surface area and surface gravity of Mars compared with the earth.)

We suggest that Mars had an adequate supply of carbon dioxide but cooled off because its recycling mechanism ran down. We think the planet once had an effective recycling system since, if it had lacked one, the weathering of rocks would have removed all the atmospheric carbon dioxide within about 10 million years. Yet the atmosphere apparently retained abundant amounts of the gas for considerably longer. The valley networks provide the clue: the ones on the old southern highlands can be dated by counting the number of meteorite craters that overlie them. On this basis workers have concluded that the networks persisted until near the end of the period of the heaviest meteoritic bombardment—about 3.8 billion years ago.

The recycling system probably removed carbon dioxide from the atmosphere by the same weathering

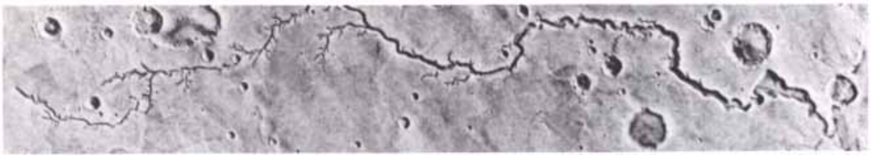


CLIMATE-MODEL CALCULATIONS indicate that the earth would have been frozen during the first part of its history if its atmospheric composition was the same as it is today. The reason is that the sun was up to 30 percent fainter in the past (*colored curve*). The top curve shows the surface temperature calculated by a "one-dimensional" (globally averaged) climate model, assuming a constant concentration of atmospheric carbon dioxide. (Many of the calculations discussed in the text are also based on a one-dimensional model.) The bottom curve shows the surface temperature of an airless earth. The shaded region between the curves represents the magnitude of the greenhouse effect. Actually carbon dioxide levels were probably higher in the past, and the surface temperature of the earth was higher than is shown here. The solar-luminosity curve is based on a calculation by Douglas O. Gough of the University of Cambridge.

process as on the earth. The mechanism that returned the gas to the environment may have been rather different, however, because a planet as small as Mars may never have developed plate tectonics. One possibility is that lava emitted by volcanoes on the Martian surface could have covered carbonate sediments and then gradually buried them to a depth at which pressure and heat would cause them to release gaseous carbon dioxide. Computer models indicate that such a process could have been sufficient to recycle carbonates for up to a billion years after the planet formed.

Mars apparently cooled down not because it received less sunlight than the earth but because it was smaller. Mars had less internal heat when it formed, and its high surface-to-volume ratio meant that it lost such heat at a higher rate. Eventually the interior became so cold that Mars could no longer free carbon dioxide from carbonate rocks. Any carbon dioxide that left the atmosphere through weathering remained locked in the ground. The Martian atmosphere became thin and the climate approached its present frigid state. If the planet had been the size of the earth, the chances are good that it would have had enough internal heat to continue recycling carbon dioxide and thus to counteract the low level of sunlight it received.

This scenario predicts that Mars now has substantial amounts of carbonate rocks buried in its crust. So far earth-based spectroscopic searches have not discovered such materials. On the other hand, James L. Gooding of NASA has recently detected small amounts of calcium carbonate in the so-called SNC (Shergotty, Nakhala and Chassigny) meteorites—which are fragments of rock that are thought to have originated at the Martian surface. The forthcoming Mars Observer Mission, scheduled for 1992, will carry out a more extensive search for carbonates and should provide important new evidence relating to our theory of how Mars froze.



SURFACE OF MARS is carved by many channels whose presence suggests that the planet was once warm enough to maintain liquid water. Typical "runoff" channels, like Nirgal Vallis (*top*), look different from rivers on the earth because their tributaries are short and sparse; they probably derive from the sapping, or leaching, of groundwater. Other, highly branched structures (*bottom*) found in ancient terrains appear to be valley networks; sapping or precipitation and runoff could have formed them. The fact that the networks are overlain by craters implies they developed before the end of the period of heavy bombardment by meteorites about 3.8 billion years ago. Certain "outflow" channels (not shown) could have formed in a cold climate by other processes.

How Venus Dried Out

Whereas Mars has a vast supply of water (albeit frozen), Venus today is almost completely dehydrated. What little water there is resides in the atmosphere as vapor or as a component of the dense sulfuric acid clouds that surround the planet. Climatologists have advanced two major theories to explain why Venus is so dry.

John S. Lewis of the University of Arizona and his colleagues have suggested that Venus never had much water—that the region of the solar nebula where Venus formed was too hot to allow for the formation of hydrated minerals. A serious problem with this theory is that it does not consider the role of gravity. According to dynamical models developed by George W. Wetherill of the Carnegie Institution of Washington, developing planets not only sweep up planetesimals that cross their orbits but also perturb the orbits of such bodies and scatter them throughout the inner solar system. During the later stages of their growth, the "proto" earth and Venus were massive enough to have actually exchanged planetesimals. Because the ones derived from the earth would have been rich in water, Venus would have received a substantial endowment of the fluid.

As this objection suggests, the alternative theory is that Venus originally had plenty of water—perhaps as much as the earth—but lost it when the life-giving substance found its way to the upper atmosphere. There sunlight tore apart the water molecules and liberated hydrogen atoms, which escaped into space. (Only water in the upper atmosphere is subject to hydrogen escape; at low altitudes hydrogen atoms, which are light, are held in the atmosphere by the drag exerted by background gases, such as carbon dioxide.)

Variations on this alternative theory differ in whether they allow water to remain fluid on the surface for any length of time. The classical explanation, the runaway-greenhouse theory, holds that Venus never retained any water on its surface. The concept of the runaway greenhouse was suggested as early as 1955 by Fred Hoyle of the University of Cambridge, but many of the details were worked out in the late 1960's by Andrew P. Ingersoll of the California Institute of Technology and one of us (Pollack).

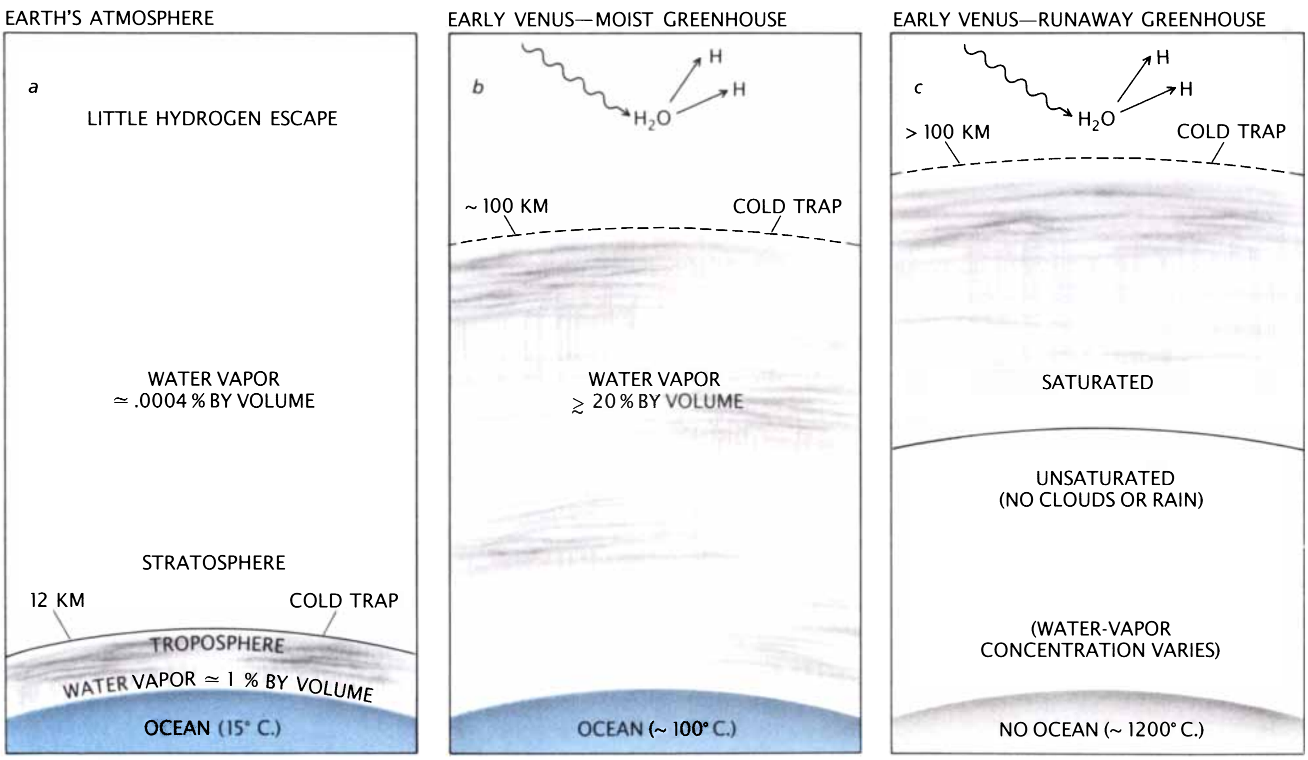
According to these workers, surface water cannot remain liquid if there is more than a critical amount of sunlight incident on a planet. If the solar flux at the orbit of Venus exceeded that critical value from the start, any water released from the interior would have vaporized instantly. At least in the lower, hotter part of the atmosphere, this vapor would not have condensed out as rain, and so no oceans would have formed.

Water would have been lost from the atmosphere because in such a hot, wet environment air would cool unusually slowly as it rose. Consequently the atmospheric "cold trap" would be pushed up to a high altitude (about 100 kilometers). The cold trap is the region where cold temperature and a high ambient pressure combine to hold the saturation point to a minimum. Normally the relative concentration of water vapor (the fraction of atmospheric volume represented by the vapor) in the cold trap is much less than the concentration in the atmosphere below it, and water condenses out instead of rising. In an elevated cold trap, however, the relative concentration of water vapor would be similar to that in the atmosphere closer to the surface. Under this condition the cold trap would allow a significant amount of water to pass into the upper reaches, where it would be subject to "photodissociation" and hydrogen escape. Such escape could potentially have eliminated the equivalent of an ocean in less than 30 million years.

In the earth's present atmosphere, in contrast, the cold trap is found at rather low altitudes (between nine and 17 kilometers), at the boundary between the troposphere and the stratosphere. When water vapor from lower altitudes rises to the cold trap, almost all of it condenses out, with the result that our stratosphere is extremely dry and little hydrogen escapes.

We calculate that the solar flux necessary to trigger a runaway greenhouse is about 1.4 times the amount of sunlight that currently impinges on the earth—if the planet in question has a fully saturated atmosphere that is free of clouds. This is approximately equal to the estimated solar flux at the orbit of Venus early in the history of the solar system, suggesting that Venus was on the brink of a runaway greenhouse. Nevertheless, if clouds were present and able to reflect a substantial fraction of the incident sunlight, it is likely that a runaway greenhouse could have been avoided on the very early Venus, allowing oceans to exist for a time.

Such oceans would not have been spared indefinitely. As an alternative to the runaway-greenhouse theory we propose that Venus once had oceans but lost them because its atmosphere was what we call a moist greenhouse: a condition in which the relative concentration of water vapor near the ground accounts for more than 20 percent or so of the volume. For a one-bar atmosphere like that of the earth, this concentration can be reached when the surface temperature rises above about 70 degrees C. (If Venus had an ocean and rain, most of its carbon dioxide would have been buried in carbonate rocks, and



TENDENCY OF WATER VAPOR to escape from the earth is minimal; the same cannot be said for early Venus. On the earth (*a*) water in the troposphere is blocked from entering the stratosphere by a cold trap: the region where cold temperature and relatively high ambient pressure combine to minimize the concentration of water vapor. When vapor reaches the trap, most of it condenses out. On early Venus the lower atmosphere, though warm by the earth's standards, may have been cool enough for water to condense and form an ocean. The sea would in time have been lost, however, to a "moist greenhouse" (*b*): a condition that arises when a high surface temperature enables water vapor to constitute more than about 20 percent of the lower atmosphere. The cold trap then moves to a high altitude and becomes inefficient at preventing water vapor from rising into the upper atmosphere. Although some vapor condenses out as rain, the steam at the top dissociates and its constituent hydrogen atoms escape into space. Venus might have been so hot that a runaway greenhouse (*c*) developed instead: all the water released by the planet turned to steam instantly, and no ocean formed. The water essentially traversed a one-way route: up and away.

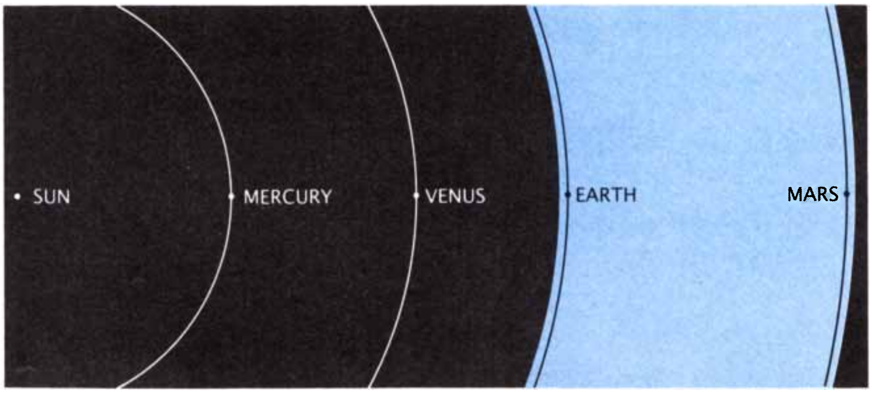
a one-bar atmosphere would have been possible.)

Our climate simulations indicate that a moist greenhouse should arise when the solar flux striking a cloud-free atmosphere is at least 1.1 times the amount of light incident on the earth. When the water-vapor concentration near the ground surpasses 20 percent, water condensation (which generates heat) warms the atmosphere significantly and, as in the runaway-greenhouse condition, causes the cold trap to rise. Water can then make its way into the upper atmosphere. An ocean could exist on a planet with a solar flux of between 1.1 and 1.4 times that of the earth, but it would be depleted by hydrogen escape within a few hundred million years.

In our view the moist-greenhouse theory does a better job than the runaway-greenhouse theory of explaining why Venus has almost no liquid water today. Because weathering would suppress the atmospheric carbon dioxide levels in a moist greenhouse, the total gas pressure of the atmosphere would be lower than in a runaway greenhouse. As a result it would take a small amount of water vapor to constitute 20 percent of the total gas volume, and so a greater fraction of the final water supply would reach the upper atmosphere. For example, if the atmosphere consisted of one bar of water vapor and one bar of carbon dioxide, the water would constitute 50 percent of the volume and much of it would escape. In contrast, if there were 99 bars of carbon dioxide, the one bar of water would constitute just 1 percent of the volume and would remain in the planet's atmosphere.

Regardless of whether the early atmosphere of Venus was in a runaway- or moist-greenhouse state, the planet would eventually have evolved to its present hot, dry condition. Once the oceans disappeared carbonate formation should have ceased, causing carbon dioxide to accumulate in the atmosphere. Consequently the planet's 93-bar atmosphere is now mostly carbon dioxide. Sulfur gases, which were initially scarce because they dissolve readily in water, also accumulated and formed the sulfuric acid clouds that are now a major feature of the Venusian environment.

It is the carbon dioxide, not the distance of Venus from the sun, that today accounts for its high surface temperature. Venus receives 1.9 times more solar radiation than the earth, but its sulfuric acid clouds reflect about 80 percent of that sunlight, so that Venus actually absorbs significantly less solar energy than the earth. Without the greenhouse effect Venus would be colder than the earth and only slightly warmer than Mars.



CONTINUOUSLY HABITABLE ZONE (*light blue*) is the region of space where a planet could theoretically maintain an earthlike climate long enough for life to proliferate. An early estimate suggested that the zone was rather narrow, extending from about .95 astronomical unit to 1.01 A.U., from just inside to just outside the earth's orbit. Newer work suggests that the outer edge may lie as far out as 1.5 A.U.—past the orbit of Mars.

The Continuously Habitable Zone

The finding that a planet with a solar flux 1.1 times the earth's would lose its water by photodissociation is consistent with Hart's calculation that the inner boundary of the continuously habitable zone lies at about .95 A.U. This agreement is somewhat coincidental, however, because we base our estimate on hydrogen-escape rates, whereas he came to his estimate by other means. Of course, a planet at the inner limit would not be habitable for long. The sun is currently increasing in luminosity by about 1 percent every 100 million years. Hence the earth itself may have difficulty maintaining its water starting about a billion years from now. This disaster may be postponed for sometime by a decrease in atmospheric carbon dioxide mediated by the carbonate-silicate cycle. Such a decline might itself prove harmful to the biota, however, because many plants would not be able to carry out photosynthesis if they received significantly less carbon dioxide than they get today. (Astute readers might note at this point that carbon dioxide levels are currently rising in the atmosphere owing to the burning of fossil fuel. In fact, such activity cannot continue for more than a few hundred years before the planet's reserves of coal and oil are eliminated. After the brief warming period carbon dioxide levels will again begin to fall.)

The outer edge of the CHZ must lie considerably farther out than Hart imagined—perhaps as far as 1.5 A.U., which would place the boundary somewhat beyond the orbit of Mars. We limit the outer boundary at this distance because it seems unlikely that a terrestrial planet could form farther out.

Much the same negative-feedback mechanism that has helped to stabilize the earth's climate for the past 4.5 billion years would presumably operate on a planet of similar size farther from the sun. The only reason Mars froze is that it was too small to continue recycling carbon dioxide. An earth-size planet at the orbit of Mars should, according to our theory, have several bars of carbon dioxide in its atmosphere and a mean surface temperature above the freezing point. This atmosphere would not be breathable by human beings, but it would be perfectly capable of supporting some form of life.

When Hart first determined that the CHZ was extraordinarily narrow, his conclusion implied that the chance of finding earthlike planets around other stars was rather slim, even if other planetary systems were themselves abundant. Our calculations point to the opposite conclusion. If other planetary systems exist, as in all likelihood they do, then there is a good probability of finding habitable planets. Whether or not any of them are in fact inhabited is, of course, an open question, but it is one that can no longer be dismissed on the assumption that the earth is climatologically unique. Perhaps on one such planet there is even an extraterrestrial version of the story of Goldilocks.

The Adaptable Opossum

The Virginia opossum can adapt quickly to a changing world. Part of its success may be due to a highly efficient reproductive strategy that includes the ability to adjust the sex ratios of its progeny

by Steven N. Austad

From the time the Spanish explorer Vicente Yáñez Pinzón and his crewmen discovered a mother opossum carrying pouch young in Brazil in February of 1500, Europeans have been moved to rhetorical excess by these New World mammals with their embryonic neonates. As is generally the case with exotic animals, opossums were initially described as if cobbled together from the parts of known species. Richard Eden, for instance, in his 1555 translation of Peter Martyr's *De Orbe Novo,* described the opossum as a "monstrous beaste with a snowte lyke a foxe, a tayle lyke a maramasette, eares lyke a batte, hands lyke a man, and feete lyke an ape, bearing her whelpes abowte with her in an owtwarde bellye much lyke unto a greate bagge or purse." Captain John Smith, leader of the Jamestown colony and the man who transliterated the Algonquian word meaning "white beast" into the English word "opassum," was somewhat more restrained, merely likening them to rats, cats and swine.

Familiarity, however, does breed contempt. Today the opossum—one of the commonest suburban mammals in the U.S.—gets little respect. Wildlife biologist Durward L. Allen, for example, derides the opossum as "a sluggish, smelly, disreputable critter without a semblance of character or self-respect." As a marsupial, moreover, it is regarded as being primitive and therefore inferior to eutherian, or placental, mammals. At a given body size marsupials have a lower body temperature, lower metabolic rate and a smaller brain than eutherians do. They also have fewer chromosomes on the average, a shell membrane around their ova and, except for the bandicoots, only a rudimentary placenta.

Yet the case for marsupial inferiority is hard to sustain, particularly as applied to American opossums. Indeed, the Virginia, or North American, opossum (*Didelphis virginiana*) is one of the ecological success stories of modern times. The species spread throughout most of North America during the late Pleistocene and it has continued its spread in historical times. Since the Colonial era, opossums have expanded their range northward in the eastern U.S. by as much as 500 miles.

In this century alone a few animals were able to establish the species on the West Coast, where it was unknown before the late 19th century. In 1906, near Los Angeles, an anonymous fur trapper captured the first wild opossum to be taken west of the Rocky Mountains. It is not clear whether this animal was descended from opossums brought from Missouri 20 years earlier by one "Uncle Billy" Rubottom or whether some other homesick Southern immigrant had also liberated a few former pets. In any case, the opossums flourished in their new home. By 1927 a wildlife biologist wrote: "The opossum is so prolific and has such a wide distribution in California that it is now too late to exterminate this dangerous species." A decade later they were thriving from Baja California to the Canadian border.

Much of the Virginia opossum's success no doubt stems from its mutualistic relationship with human beings. People provide barns, porches and a variety of other potential den sites as well as a steady supply of edible garbage. Opossums, on the other hand, provide humans with a reliable if not epicurean food source. Archaeological excavations clearly indicate that opossums were a staple of Indian diet throughout the Americas. For instance, opossums now live on a number of Caribbean islands, where there is no paleontological evidence of their presence prior to human colonization. Hence it seems likely they were brought along as a source of meat.

The range expansion of American opossums is not merely an artifact of human range expansion. Recent investigations show that opossums are extraordinarily adaptable. For instance, their small brain and dullness notwithstanding, opossums have a remarkable talent for finding food and remembering where they found it. When tested for ability to remember which of four runways was connected to a food box, opossums scored better than cats, chicks, dogs, goats, pigs, rabbits, rats and turtles, although less well than humans.

Equally surprising is the species' strong resistance to the venom of snakes in the rattlesnake subfamily Crotalinae, which also includes bushmasters, fer-de-lances, copperheads and cottonmouths. This immunity appears to be a specific adaptive response to an environmental threat, rather than an accidental by-product of their physiology, because opossums are not resistant to the bites of most Old World snakes, such as cobras and puff adders. Their immunity to crotalines has, however, enabled them not only to escape from a potential predator but also to eat the reptiles with impunity.

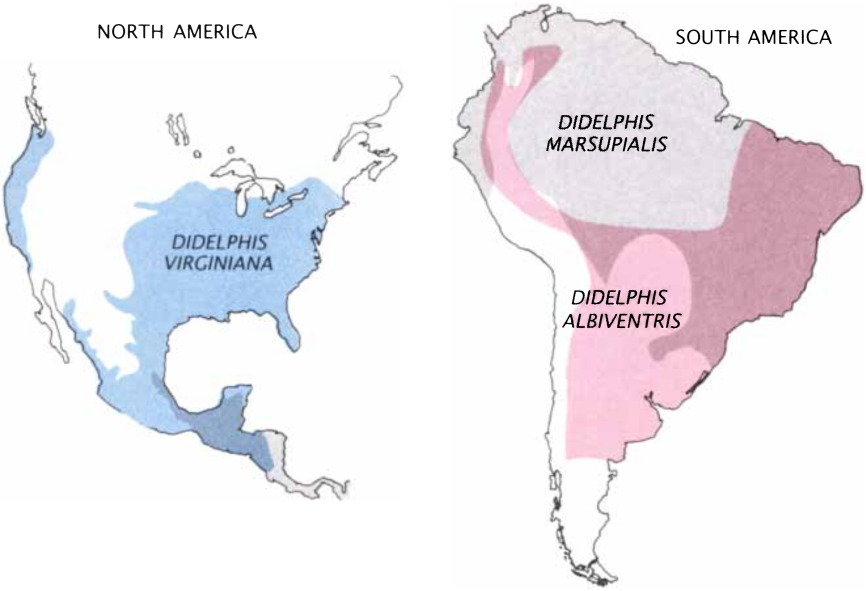
In our own studies we have found that opossums adjust the sex ratios of their offspring in a manner consistent with gaining optimal reproductive success under varying amounts of parental investment. This evidence that opossums are capable of high reproductive efficiency under changing ecological conditions could help to explain how the species was able to evolve an entire array of other sophisticated adaptations that may have allowed them to compete successfully with their more "advanced" mammalian relatives.

Marsupials and eutherians diverged some 100 million years ago, and within about 35 million years North America had a rich marsupial fauna, consisting of at least three families, five genera and 13 species. Indeed, the oldest clearly recognizable marsupial fossils are from Canada, and the prevailing (although not unanimous) opinion is that marsupials originated in the Americas, possibly in North America. For reasons not known all North American marsupials became extinct by 15 million years ago, whereas marsupials continued to radiate successfully in South America, Australia and New Guinea. Today Central and South America together have nearly 80 marsupial species, or more than 30 percent of the world's total. The American opossum family, the Didelphidae, alone accounts for about 70 species. (It is not closely related to the several Australian families of "possums.")

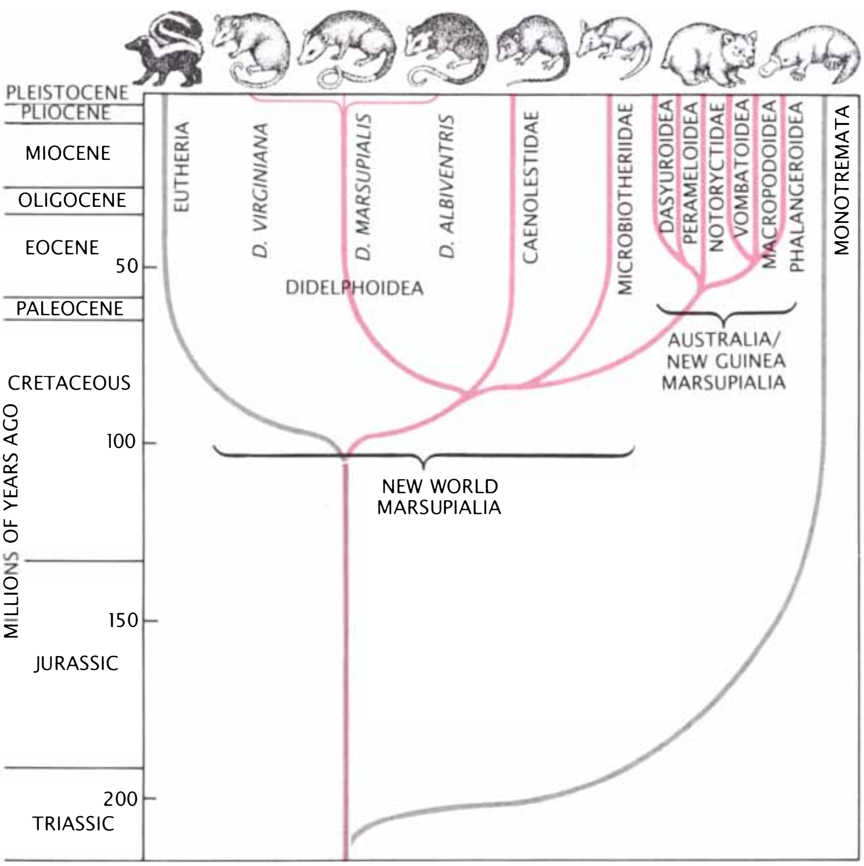
The only marsupial in the U.S. today is the Virginia opossum. It is one of three species in the genus *Didelphis,* which ranges from southern Canada to southern Argentina, making it one of the more widely distributed mammalian genera in the hemisphere and by far the most widely distributed marsupial genus in the world. Because *Didelphis* opossums bear a morphological resemblance to some of the earliest fossil marsupials, they have been erroneously called living fossils. Indeed, because of their conservative morphology, Virginia opossums are often used as research paradigms for the functional anatomy of primitive mammals.

Conservative though it may be morphologically, *Didelphis* is a recently evolved and in some ways highly specialized genus. The first positively identified *Didelphis* fossils appear about four million years ago

OPOSSUM confronts a canebrake rattlesnake with impunity. The marsupial is impervious to the venom of snakes in the rattlesnake subfamily and views the reptiles not as lethal enemies but as a food source. In parts of Texas, copperhead snakes constitute up to 6 percent of the opossum's diet. Resistance to venom may have evolved as a specific adaptive response to an environmental threat, because the species is immune to the venom of New World snakes but succumbs to that of Old World snakes.



RANGE of *Didelphis* opossums extends from southern Canada to southern Argentina, making them the most widely distributed marsupial genus in the world. The white-eared opossum (*D. albiventris*) is found chiefly in cooler climates in South America. The common opossum (*D. marsupialis*) inhabits warm, tropical woodlands. The Virginia opossum (*D. virginiana*) migrated into nontropical North America during the Pleistocene. It thrives in most habitats, excluding desert and mountains exceeding 10,000 feet. Since Colonial times its range has expanded 500 miles northward to Canada. Human settlers introduced the species to the West Coast around the turn of the century.

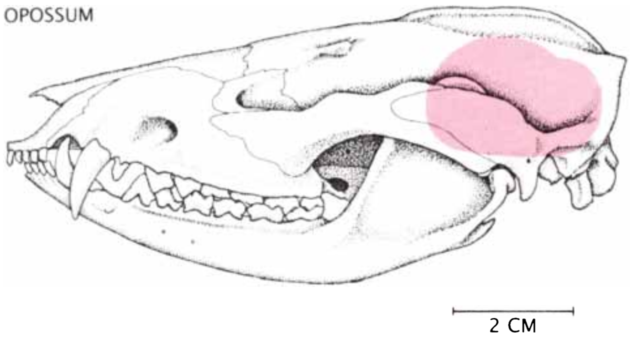


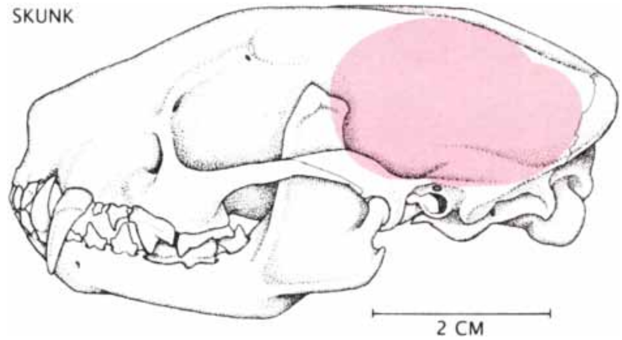
MARSUPIALS may have originated about 100 million years ago in the New World. Didelphoidea, to which the Virginia opossum belongs, is considered the oldest superfamily to survive to the present day. Morphologically *Didelphis* is remarkably similar to the earliest marsupial fossils. *D. virginiana,* which appeared only 75,000 years ago, is the sole marsupial species that has been able to range into nontropical North America.

in South America. Alfred L. Gardner of the National Museum of Natural History has convincingly argued, from chromosomal analysis, the fossil record and paleoclimatological evidence, that the Virginia opossum is the most recent of the three species in the genus, having diverged from its ancestral species, the common opossum (*Didelphis marsupialis*), in the past 75,000 years.

Gardner suggests that *D. marsupialis* colonized Mexico when land bridges reconnected South America and North America during the Pliocene, from five to two million years ago. Because they are not tolerant of arid or cold weather, however, they were not able to venture into what is now the U.S. During the Pleistocene, roughly between one million and 10,000 years ago, when repeated glaciation in North America led to major climatic fluctuation in Mexico, *D. marsupialis* spread throughout the mountains of western Mexico in warm periods but had to retreat to isolated lowland valleys in cool periods. Within one of these small isolated populations major chromosomal rearrangements occurred, which cut the group off genetically and collaterally probably helped to adapt the incipient species to cooler climates, allowing it to move northward. By 4,000 years ago the Virginia opossum had spread throughout most of the eastern U.S. Gardner also finds evidence of extreme inbreeding, which might predispose the species to be good colonizers because tolerance for inbreeding would enable a few animals to found new populations.

A further characteristic that may contribute to the opossum's ecological resilience is high reproductive efficiency. This begins even before conception with unusually efficient use of sperm. The male inseminates females with only about three million sperm, some 5 percent of which reach the site of fertilization; in comparison, a rabbit inseminates a female with about 150 million sperm, of which .01 percent reach the site of fertilization. The opossum's efficiency may be linked to sperm pairing, a feature of American marsupials that may be unique among vertebrates. In *Didelphis* two sperm conjugate side by side while they are still in the testes and then swim together through the female reproductive tract before separating in the oviduct. Some investigators think the pairing allows exceptionally good sperm transport and high





OPOSSUM'S BRAIN is small compared with a skunk's. Another primitive feature is the large number of teeth: no eutherian has more than three incisors or molars, but opossums have five incisors in the top jaw, four incisors in the bottom and four molars.

sperm survival within the reproductive tract.

After conception, female marsupials spend much less time gestating their young than eutherian females. Marsupial neonates are virtually mobile embryos, and most of their development occurs externally, often in a pouch. *Didelphis* opossums, for example, gestate their young for approximately 13 days, and each pup at birth weighs less than .2 gram, or one ten-thousandth of the body weight of the mother. In comparison, the striped skunk, a North American eutherian mammal of about the same size, gestates its young for 65 days and the kits weigh about 33 grams at birth, or more than 150 times as much as opossum neonates.

Newborn opossums are deaf and blind, and their hind limbs and tail are vestigial. Yet their forelimbs are precocially developed and equipped with deciduous claws, which assist the neonates on their trip to the pouch, a distance of four or five centimeters that each must traverse on its own without maternal help. They make their way there with a kind of swimming motion—sometimes referred to as the Australian crawl. More young are born than can be successfully raised: Virginia opossums have about 13 teats and bear some 22 young, yet their surviving litter size is usually from six to eight.

For the first 60 days the pouch young are sealed to their mother's nipples, which have swollen inside their inflexible mouths. By the age of 70 days they can grasp and release the nipples and move short distances independently. They are weaned at about 100 days, after which there is little contact between mother and offspring. All in all, then, female opossums invest about 112 days in rearing a single litter from conception to weaning. The striped skunk spends 121 days, about the same time, from conception to weaning, but whereas opossum young are completely independent after weaning, skunk young remain closely allied with their mother for another month or more. Early independence allows opossums to produce two and occasionally three litters per year compared with the single litter of skunks. Thus opossums expend a greater effort in reproduction each year.

Because development takes place externally where it can be monitored, Melvin E. Sunquist of the University of Florida and I thought opossums might be excellent animals in which to study the way animals allocate their reproductive energies. In particular, we were interested in an existing theory predicting that animals adjust the sex ratio of their offspring depending on the absolute amount of reproductive investment of which they are capable. This intuitively satisfying theory, developed by Robert L. Trivers and Dan E. Willard at Harvard University in the early 1970's, states that in mammals females capable of very large investment in each offspring should produce mainly males, and females capable of only a small investment should produce mainly females.

The underlying assumption is that the fitness of male offspring would be affected more than that of female offspring by the amount of parental investment. Because in mammals males generally have more than one mate and do not contribute to parental care, females do not usually lack mates. Consequently nearly all females are likely to reproduce, but none will produce as many offspring as the most successful males. Weak males, on the other hand, may end up with no mates or offspring. Assuming, then, that the amount of parental investment affects an animal's future reproductive fitness, males should be helped more by high investment and hurt more by low investment than females are. Hence mothers capable of high investment should produce primarily males and those capable of low investment should produce primarily females.

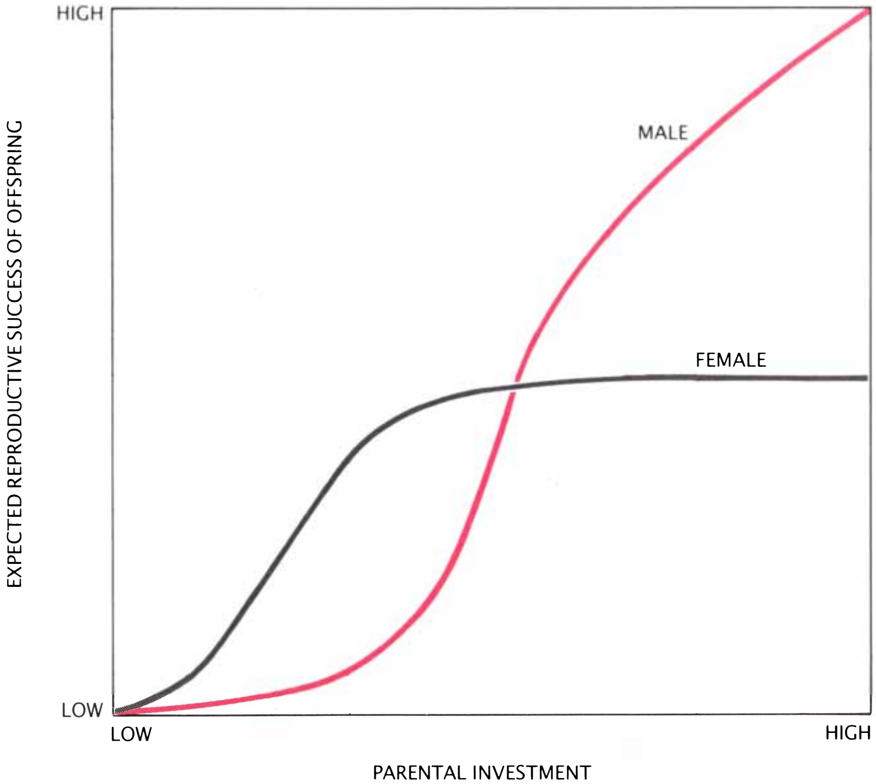
Plausible as it seems, there are reasons to be skeptical about mammals' ability to adjust offspring sex ratios. First of all, a believable mechanism is hard to imagine. The sex of the offspring is determined by whether the egg is fertilized by a sperm bearing an *X* chromosome or a *Y* chromosome, and there is no evidence that males can control the relative proportions of these sperm depending on the female they happen to be mating with. Second, domestic livestock, well-fed and cosseted, are presumably capable of much greater reproductive investment than their free-living relatives, and yet most livestock produce a balanced sex ratio. On the other hand, valid scientific theories have been developed (plate tectonics, for example) before the underlying mechanisms were discovered, and because domesticated animals have been artificially selected for such a long time, any meaning of current trends in their offspring sex ratio is not clear.

We tested the hypothesis by supplementing the food of wild female opossums, hoping we could develop specific females capable of extraordinarily high investment. Because our females wore radio collars, we could find them in their den during the day and place food near the entrance at dusk, so that they could eat soon after emerging for their nightly foraging treks. For controls we had a second group of females, which were also radio-collared and lived in the same habitat but were not fed. There were 20 animals in the test group and 18 in the control group.

Our results were clear-cut. Control

females produced a balanced sex ratio and food-supplemented females produced more males. Females in both groups bore equivalent numbers of pouch offspring. Individual pouch young from the food-supplemented group, however, were consistently larger than those from the controls, showing that the females that were fed were indeed investing more. An indication that males benefited more than females from the additional investment is the fact that our recapture rate for juveniles improved more for males than for females when we compared offspring from food-supplemented mothers with the control group.

We did no experiment to determine whether females capable of particularly little investment produced mostly female offspring, but we have some observations suggesting that is so. Most opossums in the wild live for only one reproductive season. Moreover, the few that do live into a second year show signs of advanced aging, such as cataracts, weight loss and lack of motor coordination. In captive colonies of Virginia opossums, second-year females also show atrophy of reproductive organs, decreased litter size and a higher incidence of litter failure. Those facts suggest that second-year females may not be capable of investing as much as first-year females. Our sample of second-year females is small, because they are rare, yet over the years we have managed to trap 19 such animals. Their pouch young have been female-biased to an even greater degree than the food-supplemented litters were male-biased.



SEX-RATIO HYPOTHESIS assumes that reproductive success varies more widely for males than for females in polygynous mammals: strong males will have many mates but weak males may have none, whereas nearly all females will be able to mate. Hence if the amount of parental investment affects the future reproductive fitness of offspring, mothers capable of high investment should produce more male offspring than female, and mothers capable of low investment should produce more females than males.

Our study of second-year females brings up yet another unusual feature: opossums are among the shortest-lived mammals in the world for their size. Among common opossums in Venezuela we found that of 78 radio-collared mature females, 18 percent lived to breed in a second season, 1.2 percent lived to breed in a third season and none survived to a fourth. In his study of the Virginia opossum in Florida, Sunquist found that 8.3 percent lived to a second season, 3 percent lived to a third and none lived to a fourth. Even in breeding colonies where adequate records are kept, two years is about an average life span, with from three to four years the maximum. Opossums are also unusual in that older individuals show obvious physiological signs of advanced aging in nature, something the conventional wisdom has declared rare. And yet the received wisdom may be wrong, because almost no studies of aging in wild mammal populations have been done.

Traditionally there have been two approaches to the study of aging. The physiological approach seeks to determine what immediate cellular and biochemical mechanisms underlie the aging process itself. The evolutionary approach analyzes how natural selection may account for aging and seeks evidence from the comparative aging rates of different species, or of the same species in different environments, and from the genetic correlations between aging rate and other life-history traits such as reproductive rate.

A general evolutionary theory of aging was first elaborated by the late Sir Peter Medawar in the 1940's and extended in the 1950's by George C. Williams of the State University of New York at Stony Brook. Medawar noted that even if animals did not become gradually enfeebled by age, the influence of natural selection would wane as individuals grow older because of the increasing probability of accidental death—death resulting from causes external to the animal itself, such as predators or infectious diseases. As a result deleterious mutations that act late in life will be only weakly selected against. Medawar also noted that mutations that boost early reproduction may be favored evolutionarily, even if they increase the probability of premature death.

An intuitive grasp of both aspects of this hypothesis may be had if one examines a simple example. Imagine a population of individuals that do not age. In demographic terms this means the probabilities of death and of successful reproduction are constant in the course of time. If the probability of accidental death is 10 percent per year, then after 20 years only about 12 percent of the original population will be alive; after 40 years little more than 1 percent will be alive. Any deleterious mutation that is not expressed until late in life will affect only a small fraction of the population, and any mutation that causes instant death at age 60 would have almost no selection against it. Hence there is nothing to prevent an accumulation of late-acting lethal or enfeebling mutations. If the accidental death rate is higher, mutations can accumulate that have lethal effects much earlier in life.

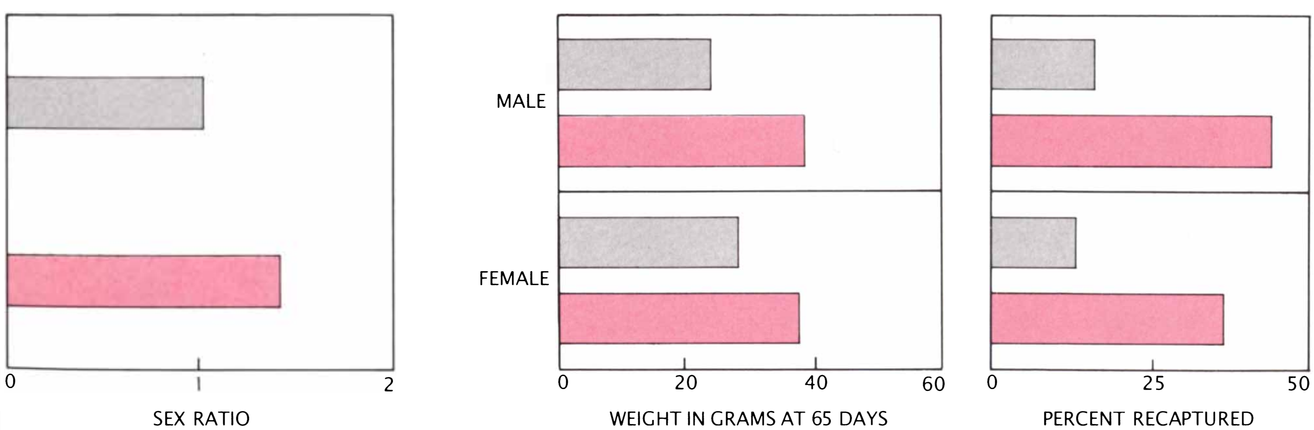
Depending on the rate of accidental mortality, there might also be an evolutionary advantage in increased early reproduction, even if there were an unavoidable survival cost associated with the increase. For instance, if the accidental mortality rate is 40 percent per year and individuals produce one offspring per year beginning at age one, then individuals will on the average produce 1.5 offspring in their lifetime. Now imagine that there is a mutation resulting in the production of two offspring per year beginning at age one, but that the cost of this excess early reproduction is a doubling of the postreproductive probability of death. These individuals would produce on the average slightly more than 1.7 offspring and so would eventually replace the original population. Again, the higher the accidental death rate, the greater the advantage of intense and early reproduction. In general, then, one would expect populations that are subject to high accidental mortality to show high reproductive effort at early ages and accelerated aging, and the converse in populations subject to low accidental mortality.

The major practical difficulty with testing the hypothesis in natural populations is to find an appropriate yardstick for aging. Longevity is not a good measure, because one cannot distinguish accidental death from senescent death. Ideally one needs a physiological measure of aging that can be applied at any age and on the same individual at different ages. One such measure may be the changing physical properties of mammalian tail-tendon fibers. Tendon is a simple tissue consisting almost entirely of bundles of parallel collagen fibers. Collagen increases in strength and decreases in solubility and flexibility in a regular manner as animals age. Moreover, the rate at which tail collagen ages responds to factors such as underfeeding and hormonal treatments that are known to affect aging in animals.

A convenient method of measuring the aging of tail collagen, developed by Harry Elden and Robert Boucek of the Howard Hughes Medical Institute in the early 1960's, is to record the breaking time of single fibers suspended by a standardized weight in a concentrated solution of urea at a standardized temperature. In laboratory rodents, breaking time has been shown to increase geometrically with the age of the animal. In the field, fiber bundles can be taken from captured animals in a quick and simple surgical procedure done under local anesthetic. The animal is then released to continue its normal existence. By sampling the same animal repeatedly one can compute the rate of change of collagen aging, which can be compared across individuals.

The two hypotheses being examined in my laboratory at Harvard are that opossums age rapidly because they expend a great effort in early reproduction and because they are subject to exceptionally high levels of accidental mortality. Our current study involves artificially reducing the reproductive effort of opossums by manipulating the size of their litters in order to see what effect this will have on collagen aging. Because prenatal reproductive investment is trivial, we can drastically influence overall reproductive effort simply by creating a range of litter sizes by transplanting young among females. We are studying the accidental death rate by comparing collagen aging, reproduction and longevity in our main study area at the Savannah River Plant in South Carolina with that of opossums living on Sapelo Island, Ga., where most of their major predators are absent and hence accidental mortality should be lower. The work is in a preliminary phase. So far we have established that the collagen technique distinguishes differences in age of as little as two months in opossums, that litter transplantation can be accomplished successfully and that the island population has been genetically isolated for at least several thousand years.

It seems ironic that the ecological success of *Didelphis* opossums has led indirectly to the low esteem in which they are generally held. That opinion might be changing, at least among scientists, as the genus helps to illuminate diverse aspects of mammalian biology. Research done to date suggests that opossums make particularly fine study animals because individuals can easily be monitored and manipulated both in the field and in the laboratory. They enable one to address broad ecological and evolutionary questions as well as to explore mechanistic details. In the future, opossums may reveal the mechanisms by which ecological forces determine sex ratios and aging rates, and physiologists may find the animals valuable for understanding such questions as the process of aging itself and the mechanisms of venom resistance.



FOOD-SUPPLEMENT EXPERIMENT showed that offspring of mothers that were given extra food (*color*) were larger and had a higher male-to-female ratio than offspring from the control group (*gray*). The recapture rate for the test group compared with the controls improved more for males than for females, implying that males derived greater benefit from high investment.

Plastics That Conduct Electricity

Cheap, durable, lightweight and versatile—plastics have a host of commendable properties, but conductivity is usually not counted among them. The next generation of plastics will flout tradition

by Richard B. Kaner and Alan G. MacDiarmid

To most people the title of this article would have seemed absurd 20 years ago, when conceptual prejudice had rigidly categorized plastics as insulators. The suggestion that a plastic could conduct as well as copper would have seemed even more ludicrous. Yet in the past few years these feats have been achieved through simple modifications of ordinary plastics. Called conducting polymers, the new materials combine the electrical properties of metals with the advantages of plastics that stirred such excitement in the 1930's and 1940's.

To make a polymer conduct electricity, small quantities of certain chemicals are incorporated into the polymer by a process called doping. The procedure for doping polymers is much simpler than the one used to dope classical semiconductors such as silicon; in fact, the techniques for making and doping conducting polymers existed long before the scientific community was ready to embrace the idea that plastics could be made to conduct. Once the potential of polymers as conductors had been demonstrated, the idea took off. In 1977 the first conducting polymer was synthesized; in 1981 the first battery with polymer electrodes was demonstrated. Last summer conducting polymers matched the conductivity of copper, and a few months ago the first rechargeable polymer battery was put on the market.

Subsequent advances suggest that polymers may be made that conduct better than copper; better, indeed, than any other material at room temperature. They may even replace copper wires in circumstances where weight is a limiting factor, as in aircraft. Conducting polymers also have interesting optical, mechanical and chemical properties that, taken together with their ability to conduct, might make them effective in novel applications where copper would not do. For instance, thin polymer layers on windows could absorb sunlight, and the degree of tinting could be controlled by means of an applied electric potential. Artificial nerves made of conducting polymers that are virtually inert might even be implanted in the body.

The discovery of conducting polymers began with an accident. In the early 1970's a graduate student in Hideki Shirakawa's laboratory at the Tokyo Institute of Technology was trying to make a polymer called polyacetylene from ordinary acetylene welding gas. The polymer, a dark powder, had first been synthesized in 1955, but no one knew much about it. Instead of a dark powder, Shirakawa's student produced a lustrous, silvery film that looked like aluminum foil but stretched like Saran wrap. Looking back over his chemical recipe, the student saw his mistake: he had added 1,000 times more catalyst than the instructions called for. What he had made was polyacetylene, all right, but in a form different from any polyacetylene before it.

When one of us (MacDiarmid) visited Shirakawa's laboratory in 1976, the search for so-called synthetic metals had just begun and in its new guise polyacetylene was an obvious target for investigation. Shirakawa subsequently spent a year at the University of Pennsylvania with MacDiarmid and his colleague Alan J. Heeger exploring the potential of the revamped polymer. The collaboration bore fruit when the three investigators tried doping the material with iodine. The flexible silver films became metallic golden sheets, and the conductivity of polyacetylene increased by more than a billion times!

Since then it has been found that about a dozen polymers and polymer derivatives undergo this fortuitous transition on doping. All consist primarily of carbon atoms and hydrogen atoms arranged in individual "monomeric repeat units" that are linked together to form the polymers [*see illustration on page 108*]. Some repeat units have nitrogen or sulfur atoms as well as carbon and hydrogen. Polyacetylene, the simplest of conducting polymers, is a continuous carbon chain in which a hydrogen atom is attached to each carbon atom. The carbon atoms are connected by alternating single and double bonds that continually trade places with one another along the carbon backbone, producing, in effect, intermediates between single and double bonds [*see illustration on page 108D*]. This conjugated configuration is typical of conducting polymers.

Although they share some structural similarities, polymers can have vastly different properties. Their sensitivity to heat and air, their solubility and the ease with which they can be molded, as well as their conductivities, vary greatly depending on the type of polymer and the method used to synthesize it. For their particular combination of traits, polyparaphenylene, polythiophene, polypyrrole and polyaniline have received a great deal of attention, but polyacetylene is still considered the prototype of conducting polymers and is the most extensively studied. It has also shown the highest conductivity of any polymer: last year workers at

COLOR CHANGES accompany the synthesis of polyacetylene, a polymer that becomes conductive on doping. The thin film is red by transmitted light when it first polymerizes on the walls of a glass reaction vessel and then turns blue when the vessel is heated. Here dry ice (*bottom*) and heating elements (*top*) hold the polyacetylene at two different temperatures to demonstrate the transition in hue. Thicker films such as the one shown on page 111 reflect light rather than transmitting it and shine like metal.

B.A.S.F. A.G. in West Germany announced that by doping a very pure form of polyacetylene they had produced a material with about a fourth the conductivity of copper by volume and twice the conductivity of copper by weight. What molecular mechanisms brought about this achievement?

The precise electronic and physical changes that accompany doping are still a matter of debate, but some fundamental aspects of the mechanism are known. Conductivity is expressed in mhos per centimeter, the inverse of the unit of resistivity, ohms per centimeter. The difference in conductivity between the insulating regime and the conducting regime is immense: good insulators such as Teflon and polystyrene have conductivities close to 10^{-18} mho per centimeter; good conductors such as copper and silver have conductivities close to 10^6 mhos per centimeter. The B.A.S.F. polyacetylene has a conductivity of approximately 147,000 mhos per centimeter.

Electricity flows by the movement of electrons; for a material to carry electricity, then, some of its electrons must be free to move around. In solids that have an extended network of atomic bonds, the electrons "move" within and between discrete energy states, called bands. Each energy band has a finite capacity for electrons, and bands can also be empty. Electrons must have a certain energy to occupy a given band. In order to move from one band to another having a higher energy, an electron needs an energetic boost.

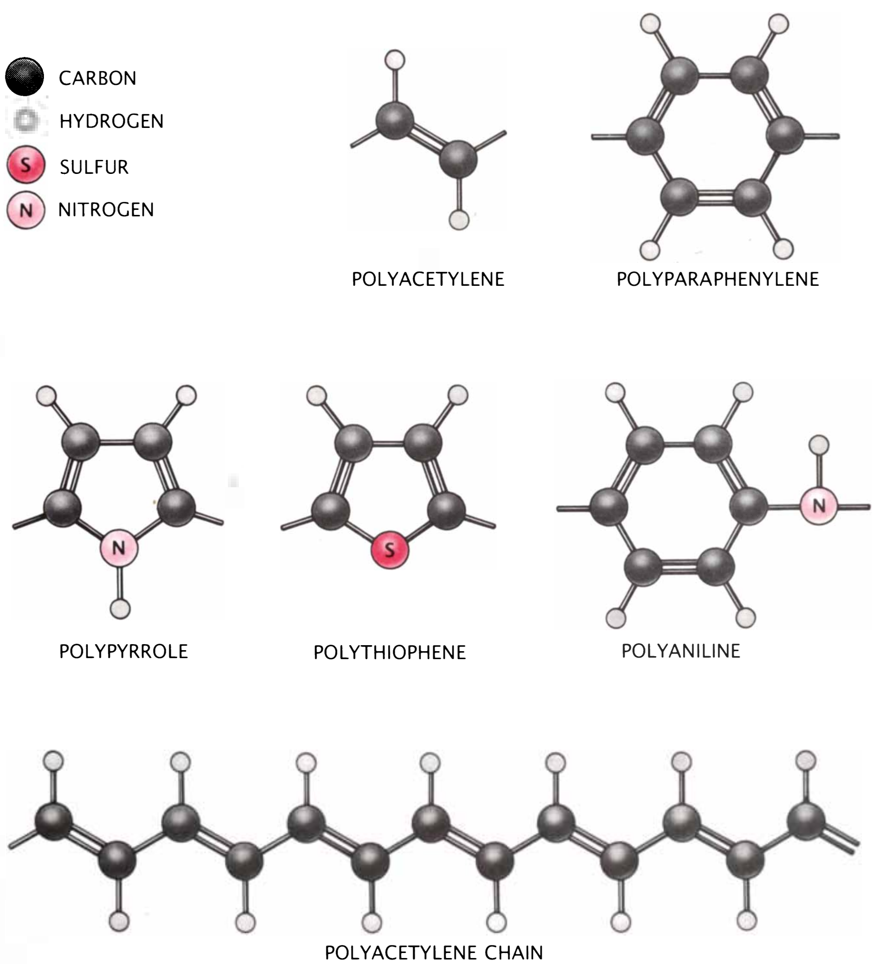
Electron movement also requires a partially filled band. Empty bands cannot carry electricity; neither can full ones. Conduction resembles a relay in which current is carried by electrons running through energy bands like racing lanes. In order to pass along the current, another "runner" must be waiting in the lane—hence the band cannot be empty. Electrons must also be able to enter the band where the runners await—hence the band cannot be full.

Metals conduct because they have partially filled energy bands. Insulators and semiconductors are materials whose energy bands are either completely full or completely empty. The highest occupied band is called the valence band and the empty band immediately above it the conduction band. A large energy gap separates the two bands in an insulator; the gap is somewhat smaller in a semiconductor. Only electrons fed bursts of energy from heat or light sources can leap across the gap.

Ordinary polymers have the electronic profile of insulators and semiconductors: their valence band is full, their conduction band is empty and the two are separated by a sizable energy gap. Doping changes that profile by taking electrons from the valence band (called *p*-doping for positive doping, since the molecule acquires a positive charge) or by adding electrons to the conduction band (called *n*-doping because the molecule acquires a negative charge). Thus in principle the polymer becomes a conductor because either the valence or the conduction band ends up being partially filled.

Actually the transition is a little more complicated. The charge conferred on a polymer backbone by doping causes a slight but important change in the position of its atoms. This change prompts the formation of one of three types of charge "islands" called solitons, polarons and bipolarons. The islands form near dopant ions and at high doping levels; when there are about 15 carbon atoms for every dopant molecule, the islands start to overlap. Adjoining islands create new energy bands, between and even overlapping the valence and conduction bands, through which electrons can freely flow. And so solitons, polarons and bipolarons are responsible for making polymers conduct.

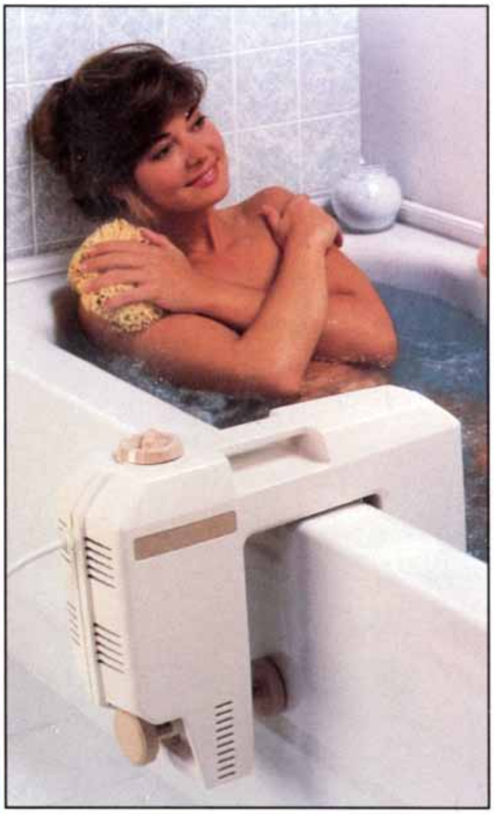
The exotic energy bands in conducting polymers do not occur in doped semiconductors. They may, however, be involved in the conduction mechanism of the new high-temperature superconductors. Indeed, the discovery of solitons in doped polyacetylene has captured the interest of theoretical physicists world-



MONOMERIC REPEAT UNITS join in chains to make up the polymers designated below each unit. The salient feature of these molecules is their alternating single and double bonds, which create an electronic configuration that is susceptible to doping. At the top are repeat units for the five most popular polymers; at the bottom is an undoped chain of polyacetylene. In their undoped form these polymers do not conduct electricity.

Shopping Time. Today, it's nearly non-existent. The Lifestyle Resource brings the store to you where you can shop at leisure. When you buy from us, you have the pleasure of knowing you are choosing among the best products of their kind. But your satisfaction is our primary concern. If you purchase a product from us that doesn't meet your expectations, just return it for a refund.

YOUR OWN WONDERFUL WHIRLPOOL



If you've ever been to a spa or health club and experienced the incredible luxury of a whirlpool bath, then listen closely: you can enjoy the same soothing relaxation in the privacy of your own home, for just a fraction of what a spa visit or a health club membership would cost. The Dazey TurboSpa attaches to any tub — large or small, round or square — and provides the same relaxing water massage previously available only in expensive built-in units. Just slightly larger than an attache case, the lightweight TurboSpa quickly and securely clamps to the side of the tub and plugs into any wall outlet. A twist of the 0-60 minute timer dial starts the powerful ¼-HP motor, which can be regulated through 10 power levels to provide anything from a light end-of-day massage, to a stimulating post-workout whirlpool that moves an amazing 42 gallons per minute. Relieve tension or aching muscles — the high-output venturi discharge tube can be rotated 180° to direct the water massage where needed. TurboSpa is encased in high-impact plastic, and all electrical components are double-insulated and located outside the tub. Leave it permanently mounted, or store it easily in a closet or under the sink. TurboSpa is built to last, and offers more power and versatility than anything else at or near its price. **$129.95** #1910.

EXPERIENCE NATURAL STATES...THE VIDEO

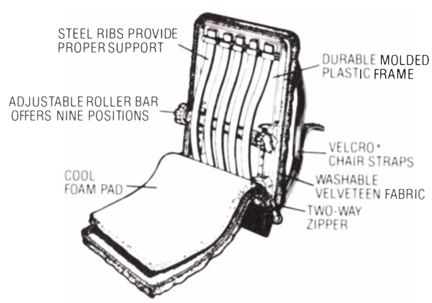
At last — the real thing in relaxation videos. No more inane fishtanks or fireplaces: Miramar's "Natural States" cassette presents a 45 minute trip through the Pacific Northwest, from California's Big Sur to Washington's Cascade Mountains. This is NOT a travelogue — not a word is spoken as the unique "steadicam" camera glides through rainforests, past verdant waterfalls, over and around mist-shrouded peaks. This unusual technique, mated to superb aerial shots, produces something very nearly like an out-of-body experience. Incredibly crisp footage enhanced by an award winning score that was specially composed "note-to-frame," rather than just layered on in post-production. This is the first nature/music video to "go gold" in sales, and for good reason: its breathtaking beauty will bring tranquility to the worst case of stress and burnout. Just slip "Natural States" into your VCR; in 45 minutes you'll be human again! Hi-fi Stereo with Dolby Surround plays mono on regular machines. **$29.95** #2010.

IMPROVED DIVERS WATCH

Our improved divers watch is a highly accurate instrument that is handsome enough for daily wear yet rugged enough for depths of 150′. It has luminous hands and an LCD display. The LCD display shows time continuously in 12 hr. or 24 hr. mode and has an illuminating lamp for easy reading at night. Push button to display day and date. Stopwatch is accurate to 1/100 second. Alarm and hourly time signal. Rotating bezel, mineral glass crystal, highly scratch resistant Teflon coated adjustable wristband. **$39.95** #1051 Black; #1061 Stainless and gold-plated.

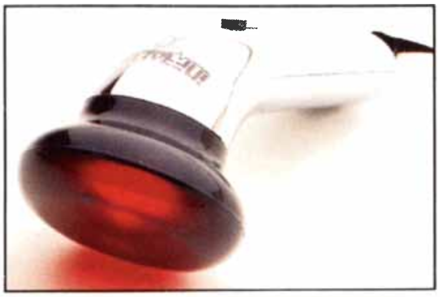
GIVE YOUR BACK A BREAK

Lower back pain is one of the commonest of human ailments. If you suffer from it, or if you want to avoid it, *you need this*— The Back Machine. Developed in cooperation with leading orthopedic physicians, and in wide use by hospitals and physical therapists, this remarkable 12″ x 17″ device—no bigger than a cushion—compensates for stress, fatigue, bad posture or poorly designed chairs. Unlike other units that give support in only one place, The Back Machine adjusts to nine positions to conform to your own unique spinal geometry: just unzip the washable cover and rotate the adjustable roller until the lumbar curve is precisely where you want it. Now, your spinal column receives the support it requires and lower back stress is minimized. Perfect for home, office, the car or plane (especially those long trips) it's light weight and portable. Durable steel and plastic construction, with velcro straps to attach to any chair. **$39.95** #1870.



THE NEW INFRALUX

Pain emanating from inflamed joints and muscles can be excruciating. Just ask sufferers of bursitis, neuralgia, or arthritis. Infrared heat is the natural remedy preferred by doctors and therapists to dispel such nerve-racking pain. Recently redesigned, the Infralux infrared therapeutic heating unit for personal use now is contoured, angled for reaching back and neck. The heating surface is tripled and the replaceable 10-watt bulb gives 45% more heat. Hi/low settings. Plugs into household outlet. UL listed. Unit with case **$39.95** #1140.



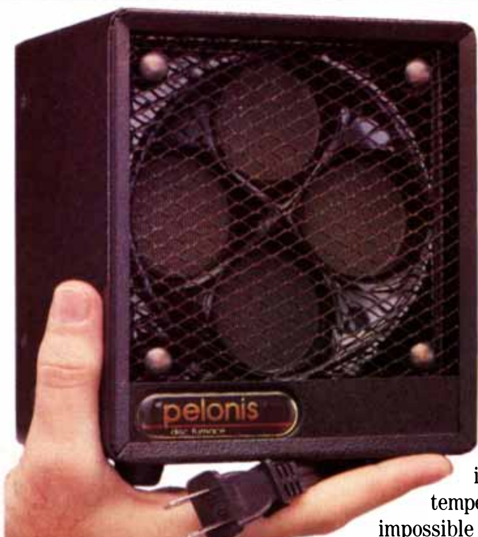
YOUR SMILE IS BEAUTIFUL

But you're only fooling yourself if you think it's proof of healthy teeth and gums. The unseen enemy is a natural mouth substance deposited on the teeth - plaque. Unchecked, plaque develops into rock-hard tartar behind which decay breaks down the tooth enamel and invades the soft, sensitive substructure. You have to remove the plaque before it hardens to prevent the process. Until recently there was no easy solution to this problem, but now there's Interplak, a precision-engineered, university-tested, three-speed plaque removal instrument. Its 10 independently driven bristle tufts rotate around their own axes: each tuft spins in the direction opposite to the spin of its neighbor at 4,200 rpm. The bristles are soft and rounded, adhering to dentistry recommendations, so that virtually every particle of plaque—*98% of it*—is removed gently and without irritation. Interplak even reaches between teeth and around orthodontic braces. Children think it's fun, so they're eager to brush. Interplak is cordless and easily stores in a free-standing or wall-mounted recharger/holder. Individual color-coded brush heads are available. The Interplak Home Plaque Removal Instrument with charger and two brush heads. **$99.95** #1110. Two additional brush heads **$29.95** #1111.

Hand brushing's motion misses plaque.

The counter-rotational movement cleans plaque.

THE SIX INCH, SIX POUND FURNACE



Imagine heating a 17′x 20′ room for 24 hours for 68 cents! In tests conducted by the National Certified Testing Laboratories, the Pelonis Disc Furnace raised the temperature of a 17′x 20′ room from 50°F to 70°F and maintained the temperature at that level for 24 hours. The costs were based on the national average for electricity of 7 cents per kilowatt hour. The secret of the revolutionary Pelonis Disc Furnace is a scientific breakthrough: barium titanate honeycomb disc heating elements. Let's look more closely at some of the unique features of this remarkable product. *Safety:* despite its enormous heat output of 5,200 BTU's, the tiny Pelonis is incredibly safe. Utilizing special ceramic low temperature heating elements, it is virtually impossible for the heating elements to ignite even something as flammable as tissue paper. There are no combustible fuels, flame or red-hot coils and a safety grille gives added protection. Exceptionally safe and great for a child's room. *Efficiency:* the honeycomb pattern in the disc elements produces an extremely large heating surface, and the quiet turbo fan drives 100% of the air through. No other heater can do that. As a result the Pelonis does its work much more efficiently than an ordinary electric heater. *Comfort:* the Pelonis is the only heater with a true thermostat that allows you to dial in the temperature you want. The patented thermistor circuitry then holds that temperature, without the heating peaks and valleys that can result from cycling. *Durability:* the Pelonis is built like a tank, with a heavy-gauge metal case, industrial-grade components, and a meticulous attention to detail. It carries a 5-year limited warranty. *Versatility:* this six pound, six inch cube will go anywhere. The Pelonis allows you to lower the heat for the entire house and selectively heat only those areas you occupy. The Pelonis Disc Furnace was introduced in Canada (where they know a lot about cold) in 1984; last year it outsold all other electric heaters combined! We are pleased to introduce this latest technology in electronic heating. Includes a built-in handle and washable air filter. Standard 110 volts AC, UL listed. **$179.95** #1810.

ZONE OF CALM

Noise produces stress in the human organism. Today the volume of civilization nearly everywhere seriously interferes with our abilities to relax, read, sleep, concentrate or otherwise function at optimum efficiency. But you needn't be victimized by noise pollution. The Marsona Sound Conditioner electrically synthesizes a variety of pleasing natural sounds that help mask out and reduce the annoyance of unwanted noise. You can simulate the sounds of ocean surf, summer rain, mountain waterfalls. You control not only the volume but also wave pattern, wave or rain rhythm, the seeming nearness or distance of the source. The Marsona is the finest instrument of its kind on the market, solid-state electronics designed for continuous use (electrical draw only 5 watts). UL-listed. **$139.95** #1860.

SHOP BY MAIL **THE LIFESTYLE RESOURCE™** SHOP BY PHONE

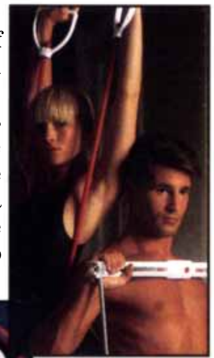
THE COLD BUSTER

Now a major scientific breakthrough—The Viralizer® System—prepares you for the onslaught of the cold season. It's the newest development of a concept pioneered at the Pasteur Institute in Paris. The cause of the common cold is the Rhinovirus family which lives and multiplies in the nose and throat, but cannot thrive in temperatures over 110° F.

The Viralizer is designed to deliver a gentle, controlled heat which penetrates the nose and throat, creating a hostile environment for cold viruses. After a pleasant heat treatment, the Vira-Spray dispenses either of two mild, over-the-counter, medicated sprays. Vira-Spray I is an analgesic, anti-bacterial spray. Vira-Spray II is a decongestant. These therapeutic sprays further discourage the stubborn cold germs so you're less likely to be re-infected or spread your cold to others. The Viralizer can produce effective relief by using it for only 3 or 4 minutes, several times a day. Proven in clinical tests 90% effective on eliminating the symptoms of upper respiratory infection in 24 hours or less, the Viralizer works without pills. Viralizer, is safe for children and adults, has been tested and recommended by doctors. The complete Viralizer® System includes 1 electric Viralizer with Vira-Spray I and Vira-Spray II plus a 3-pak refill of medicated sprays. **$39.95** #1690.

THE LIFELINE GYM

The Lifeline gym is the most space efficient, time efficient home fitness device on the market today. You can simulate just about any of the exercises done on the expensive machines found in health spas and gymnasiums. The reason is latex—a unique stretchable latex resistance cable, similar to the type used to stop planes on aircraft carriers. With this latex cable, resistance increases with movement—your muscles are challenged through their full range of motion. These variable resistance exercises provide an excellent workout for a beginner as well as a challenge for a pro. The Lifeline Gym is so successful it is used by the Chicago Bears, The U.S. Ice Hockey team, and the UCLA track team to name a few. Adjustable for individual strength levels and for different body shaping and toning. The gym includes a lifting bar, resistance cable, stirrups, exercise belt, door attachment and a comprehensive fitness guide that outlines over 25 exercises. Everything packs into a compact carry case, so you can enjoy the benefits of a complete workout anywhere. **$49.95** #1840 Ladies or #1850 Men's.



The Lifeline Gym is so successful it is used by the Chicago Bears, The U.S. Ice Hockey team, U.S. Women's Volleyball Team, and the UCLA track team.

8K POCKET COMPUTER

Introducing the world's first double line 8K pocket computer, no larger than a credit card. With the revolutionary 8,000 character Platinum Business Card you can create a database: store up to 500 names, addresses and phone numbers, easily add, delete or edit information, program 6 currency exchange rates into the memory and also have a 4 function calculator. Then at the touch of a button, view it on the two line, easy to read, 40 character (the biggest of its kind) screen. Maintain a travel expense record, enter messages or appointments — then retrieve data quickly with the automatic alphabetical sort. Security code safeguards confidential data. Battery with memory backup included. **$49.95** #1820.

JEWELRY TAKES A SHINE

Brighten up your jewelry and restore the luster it had when you first wore it. Without the bother of taking it to a jeweler, or the difficulty of trying to clean it yourself. You'll welcome this remarkable, quick and easy process made possible by the science of ultrasonics. High-frequency sound vibrates ever so gently, creating unseen waves which act on a specially formulated cleaning solution. Makes jewelry come clean in minutes with no effort on your part. Everything gleams, sparkling new again—precious and semi-precious metals and stones. Even works well on dentures, using regular denture cleanser tablets. Complete twin tank system includes 6 oz. bottle cleaner solution, rinse tray, brush, dust cover. UL listed. **$36.95** #1630.

WHEN YOU'RE IN A HURRY

When you're running short of time, ask one of our phone operators about UPS Blue Label delivery. In most cases, we can get your shipment to you in three or four working days.

CALL TOLL-FREE 800-872-5200

THE LIFESTYLE RESOURCE™ — CREDIT CARD ORDERS CALL TOLL-FREE 24 HOURS **800-872-5200**

DEPT. SFAB08; 921 EASTWIND DR. SUITE 114: WESTERVILLE, OH 43081

ITEM NO	QTY.	DESCRIPTION	ITEM PRICE	TOTAL PRICE
			SUB TOTAL	
			SHIPPING (see table at left)	
			TOTAL	

SEND TO (PLEASE PRINT)

ADDRESS

CITY STATE ZIP

☐ CHECK OR MONEY ORDER ☐ MASTERCARD ☐ VISA ☐ AMEX SIGNATURE

ACCT # EXP. DATE

Shipping Charge covers UPS. handling and insurance for guaranteed delivery.

Up to $20 .... $ 3.95	$50.01 to $60 .... $ 7.95
$20.01 to $30 .... $ 4.95	$60.01 to $70 .... $ 8.95
$30.01 to $40 .... $ 5.95	$70.01 to $100 ... $10.95
$40.01 to $50 .... $ 6.95	Over $100 .... $12.95

UPS 2ND DAY AIR

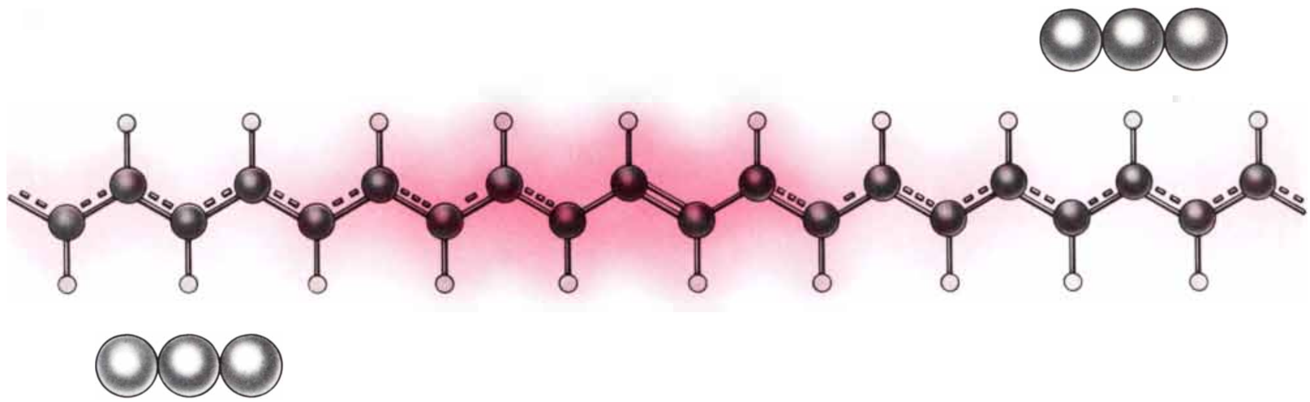
UPS Second Day available for an additional $7.50 per order.

ORDER WITH CONFIDENCE

- *We ship most orders within 48 hours.*
- *Credit Card orders billed only upon shipment.*
- *Satisfaction guaranteed or return for prompt refund.*

THE MEDIA DEVELOPMENT GROUP

©1988 G27 ADVERTISEMENT



P-DOPED POLYACETYLENE CHAIN has a delocalized positive charge that enables it to conduct like a metal. The electron density (*color*) is smeared along the polymer backbone in an internally structured cloud that diminishes near the negative dopant ion. The distinction between double and single bonds also becomes blurred near the dopant, as is indicated by broken lines.

wide because the conducting polymer provides an excellent system for studying this unique phenomenon.

The relation of the dopant molecule to the "dopee" is also different for classical semiconductors and polymers. The doping of silicon, for example, involves actually replacing some of the silicon atoms with atoms that have either more or fewer electrons. In the doping of conducting polymers the dopant molecules never replace any of the polymer's atoms; they simply act as associates that remove or contribute electrons to the polymer's energy bands. At high doping levels, when the soliton, polaron or bipolaron islands merge, the charge created on the polymer is more delocalized, smeared out across many carbon atoms.

While a great deal of uncertainty still surrounds the details of intrachain electron transport, investigators are also at a loss to explain exactly how electrons pass between chains. Yet pass they must, because a given chain does not traverse the breadth of a piece of polymer film. Some kind of interchain hopping may occur; alternatively, in the case of polyacetylene doped at intermediate levels, charge may be carried by a mobile, "liquid" soliton lattice that bathes proximate polymer chains.

And no one theory can explain why conducting polymers have such a wide range of conductivities. Nevertheless, investigators have an empirical familiarity with the polymers that has helped them to disclose some of the factors that influence conductivity. The doping percent is the most obvious determinant, and it can be precisely controlled. Another factor is the orientation of the polymer chains in the material. Chain structure gives rise to anisotropy, or asymmetry; that is, a chain conducts better in one direction than in another. In doped polyacetylene, for example, conductivity can be as much as 1,000 times better along the chain axis than it is across the molecule.

The conductivity of doped polymer films can be maximized by aligning the chains so that they run parallel. In fact, this alignment is partially realized in the B.A.S.F. polyacetylene. After the polymer has been synthesized and before it is doped, the polyacetylene sheets are stretched to orient the fibers [*see illustration on opposite page*].

Purity, too, affects conductivity. Impurities can interfere with electronic mobility by producing defects that block the passage of electrons through polymer chains. Many workers attribute the performance of the B.A.S.F. polymer to the process by which it is synthesized, which produces a remarkably homogeneous and pure polymer.

Polymers are synthesized and doped by either chemical or electrochemical processes. The standard process for making polyacetylene is a chemical one, not too different from that carried out by Shirakawa's graduate student. The inside of a glass vessel is coated with a chemical catalyst that encourages polymerization and then acetylene gas is released into the vessel. A coherent film begins to grow on the glass; within five minutes a layer of undoped polyacetylene as thick as a piece of paper coats the vessel. After it has been washed, the film can be peeled from the sides of the vessel and then, for example, *p*-doped by treatment with an iodine solution or *n*-doped with a solution of sodium metal in mercury.

With electrochemical techniques, synthesis and doping occur simultaneously. Two metal electrodes are dipped in a solution containing the dopant ion and the monomer that will constitute the polymer. In a typical *p*-doping synthesis an electric potential applied to the circuit removes electrons from the monomers adjacent to the positive electrode. This prompts the monomers to polymerize on the surface of the electrode.

During polymerization electrons are still being pulled from the nascent polymer, which becomes positively charged and so attracts negative ions—the dopant ions—from the solution. A film of the doped polymer accumulates on the positive electrode and can then be stripped off. The percent of doping depends on the size of the positive charge the polymer bears, which in turn depends on the final potential applied to the polymer. By adjusting that potential one can control the amount of dopant present in a polymer and hence its conductivity.

Polyacetylene as well as many other polymers can be *n*- or *p*-doped electrochemically after they have been chemically synthesized by attaching strips of the polymer to either the positive or the negative electrode, immersed in a solution containing the dopant ion. With steady improvements in structure, purity and doping methods the conductivity of polymers could someday surpass that of copper by volume as well as by weight.

Of course, conductivity is only one of the criteria that determine whether a material is practical. When conducting polymers were first synthesized, it was thought some of their chemical and physical properties might restrict technological applications. The conductors were insoluble and could not be melted, and so they could not be cast in molds like ordinary polymers. Many of them also decomposed when they became exposed to air, which greatly limited potential applications.

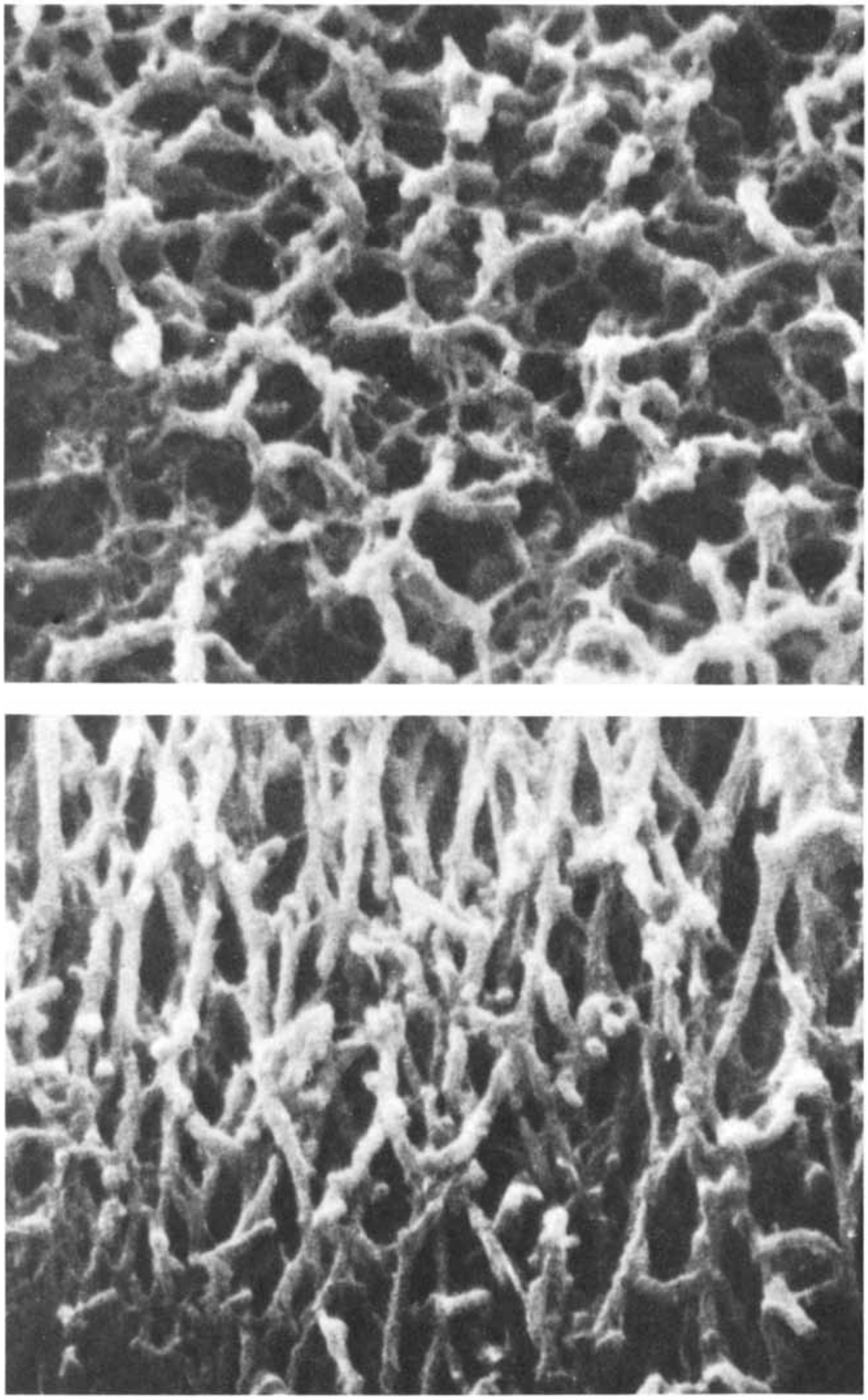
In the past two years the first of these problems has been surmounted by attaching to the carbon backbone other chemical groups that alter the properties of the molecule. For example, work at Allied-Signal, Inc., has shown that polythiophene can be rendered soluble in common solvents by replacing a hydrogen atom with the larger hydrocarbon groups such as the four-carbon butyl group. Furthermore, when a chain of 12 carbons is attached to each monomeric repeat unit, the polymer can be melted and spun into fibers that conduct in the metallic regime when doped with iodine.

With these improvements the prospective uses for conducting polymers have multiplied. The most imminent and publicized application is the rechargeable battery. In the early 1980's MacDiarmid and Heeger demonstrated that *p*- and *n*-doped polyacetylene could serve as the electrodes of a prototype rechargeable battery. The polymers were heralded as the invention that would make the electric car a reality because lightweight polymer batteries would not weigh it down to the extent that heavy lead-acid batteries would.

Although the impetus to develop electric cars has dwindled, the appeal of rechargeable batteries with polymer electrodes has not. It is thought polymer electrodes would last much longer than metal ones because the ions involved in the delivery and storage of charge come from the solution rather than from the electrodes themselves. Thus polymer electrodes would be spared the mechanical wear brought on by the dissolving and redeposition of electrode material that occurs during charge-discharge cycles in ordinary batteries. Furthermore, polymer batteries, unlike nickel-cadmium and lead-acid batteries, do not contain toxic materials and would therefore minimize disposal problems.

The Bridgestone Corporation and Seiko Electronics Parts have jointly developed a coin-type rechargeable battery with a polyaniline electrode that went on sale in Japan a few months ago. The other electrode is made of lithium metal. The batteries are said to have a capacity three times that of existing rechargeable lithium batteries of the same type and a voltage two to three times as high as nickel-cadmium batteries or one and a half times as high as the lead-acid batteries in cars. Their self-discharge rate, which governs the "shelf life" of a battery, is also significantly less than those of nickel-cadmium and lead-acid batteries.

In West Germany, B.A.S.F. and Varta Batterie A.G. are testing a flexible rechargeable battery with one electrode made of lithium metal film and the other of polypyrrole film. One of the experimental models is a postcard-size cell about a sixth of an inch thick that delivers three volts and can be bent like cardboard. Allied-Signal is also working on a re-



STRETCHING POLYMER FILMS aligns the fibers and improves conductivity in the direction of stretching. These electron micrographs show polyacetylene strands before (*top*) and after (*bottom*) stretching; each fiber consists of about 1,500 polymer chains.

chargeable battery with a conducting polymer electrode. The company has not disclosed all the details of its operation, but it appears the battery may store almost twice as much electric energy on a weight basis as a comparable nickel-cadmium cell.

The progress with polymer electrodes is quite encouraging, but the potential of conducting polymers goes beyond batteries. Allied-Signal already has several projects under way to exploit the application of conducting polymers in sensing devices. There are several strategies it could pursue. Some doped polymers decompose when they become warm, with a concomitant loss of conductivity. Hooked up to an ohmmeter, such polymers could monitor the exposure of drugs during shipment to temperatures that would threaten their efficacy. Conducting polymers have already been incorporated in sensors that reveal whether frozen food has been thawed and subsequently refrozen during handling. In a similar manner the presence of toxic substances that act as dopants for an undoped polymer can be gauged. The polymers could also act as radiation sensors if they were placed in gases that become active dopants on exposure to radiation.

Research is being carried out in several countries on the use of polythiophene and polyaniline in electrochromic display devices, which change color on the application of an electric potential. Thin films of polythiophene, for example, are red in the doped state and deep blue in the undoped state. Applications for the displays would include timetables in airports and train stations and a novel medium for indoor advertising, as well as calculators, computers, clocks and just about any other piece of equipment that currently utilizes a liquid crystal display (LCD). Although they could be made in a variety of colors, it is hard to tell now whether electrochromic displays would have distinct advantages over LCD's. In order for such devices to be practical both the switching time and the operating life of the polymers must be considerably improved.

The same mechanism by which conducting polymers change color in an electrochromic display could also be exploited in ordinary window glass. Very thin polymer layers, embedded in a colorless solid electrolyte and sandwiched between two layers of glass, could tint a window when an electric potential was ap-

POLYACETYLENE SHEET is stripped from the walls of a reaction vessel after synthesis. As flexible as paper, the freestanding film must be doped to its conducting form.

plied. The degree of tinting could be set by the size of the potential. Furthermore, because conducting polymers can absorb sunlight, they have attracted the attention of investigators working on solar-energy cells. To date, however, the conversion efficiencies have been disappointing.

Conducting polymers also absorb electromagnetic energy at low frequencies and could therefore be employed as electromagnetic shields to stop radiation leakage from computer terminals. Currently metal- or carbon-filled plastics serve this purpose, but the processability and higher conductivity of homogeneously doped polymers give them special advantages. The high microwave-loss properties reported recently for polyaniline hold promise for this particular type of application.

Doped polymers have been tested as the conductive pathways in novel printed circuit boards, in transistors and as substitutes for conventional semiconducting junctions, but their performance so far falls short of that of current materials. Nonetheless, refinements in polymer synthesis to yield molecules with better alignment and fewer defects, as well as advances in doping techniques, will undoubtedly lead to marked improvements in performance. Because of their unusual optical properties, conducting polymers are also being considered for use in the imminent generation of optical computers.

The human body is another "device" in which conducting polymers might someday play a part. Because they are inert and stable, some polymers have been considered for neural prostheses—artificial nerves. Polypyrrole in particular is thought to be nontoxic and can reliably deliver an appropriate electric charge. The dopant ion here might be heparin, a chemical that inhibits the clotting of blood and is known to function quite adequately as a dopant in polypyrrole. Alternatively, polymers could act as internal drug-delivery systems, planted inside the body and doped with molecules that double as drugs. The drug would be released when the polymer was transformed to its neutral state by a programmed application of an electric potential.

In many ways the status of conducting polymers in the mid-1980's is similar to that of conventional polymers 50 years ago. Although conventional polymers were synthesized and studied in laboratories around the world, they did not become technologically useful substances until they had been subjected to chemical modifications that took years to develop. Likewise, the chemical and physical properties of conducting polymers must be fine-tuned to each application if the products are to be economically successful. Regardless of the practical applications that might be found for conducting polymers, they will certainly challenge basic research in the years to come with new and unexpected phenomena. Only time will tell whether the impact of these novel plastic conductors will equal that of their insulating relatives.

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, Utah 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, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/88. $1.25 + 0.00.

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

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 953, Farmingdale, N.Y. 11737. Telephone inquiries: 1–800–227–5782. The date of the last issue on subscriptions appears at the 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 ____________________

Ramanujan and Pi

Some 75 years ago an Indian mathematical genius developed ways of calculating pi with extraordinary efficiency. His approach is now incorporated in computer algorithms yielding millions of digits of pi

by Jonathan M. Borwein and Peter B. Borwein

Pi, the ratio of any circle's circumference to its diameter, was computed in 1987 to an unprecedented level of accuracy: more than 100 million decimal places. Last year also marked the centenary of the birth of Srinivasa Ramanujan, an enigmatic Indian mathematical genius who spent much of his short life in isolation and poor health. The two events are in fact closely linked, because the basic approach underlying the most recent computations of pi was anticipated by Ramanujan, although its implementation had to await the formulation of efficient algorithms (by various workers including us), modern supercomputers and new ways to multiply numbers.

Aside from providing an arena in which to set records of a kind, the quest to calculate the number to millions of decimal places may seem rather pointless. Thirty-nine places of pi suffice for computing the circumference of a circle girdling the known universe with an error no greater than the radius of a hydrogen atom. It is hard to imagine physical situations requiring more digits. Why are mathematicians and computer scientists not satisfied with, say, the first 50 digits of pi?

Several answers can be given. One is that the calculation of pi has become something of a benchmark computation: it serves as a measure of the sophistication and reliability of the computers that carry it out. In addition, the pursuit of ever more accurate values of pi leads mathematicians to intriguing and unexpected niches of number theory. Another and more ingenuous motivation is simply "because it's there." In fact, pi has been a fixture of mathematical culture for more than two and a half millenniums.

Furthermore, there is always the chance that such computations will shed light on some of the riddles surrounding pi, a universal constant that is not particularly well understood, in spite of its relatively elementary nature. For example, although it has been proved that pi cannot ever be exactly evaluated by subjecting positive integers to any combination of adding, subtracting, multiplying, dividing or extracting roots, no one has succeeded in proving that the digits of pi follow a random distribution (such that each number from 0 to 9 appears with equal frequency). It is possible, albeit highly unlikely, that after a while all the remaining digits of pi are 0's and 1's or exhibit some other regularity. Moreover, pi turns up in all kinds of unexpected places that have nothing to do with circles. If a number is picked at random from the set of integers, for instance, the probability that it will have no repeated prime divisors is six divided by the square of pi. No different from other eminent mathematicians, Ramanujan was prey to the fascinations of the number.

The ingredients of the recent approaches to calculating pi are among the mathematical treasures unearthed by renewed interest in Ramanujan's work. Much of what he did, however, is still inaccessible to investigators. The body of his work is contained in his "Notebooks," which are personal records written in his own nomenclature. To make matters more frustrating for mathematicians who have studied the "Notebooks," Ramanujan generally did not include formal proofs for his theorems. The task of deciphering and editing the "Notebooks" is only now nearing completion, by Bruce C. Berndt of the University of Illinois at Urbana-Champaign.

To our knowledge no mathematical redaction of this scope or difficulty has ever been attempted. The effort is certainly worthwhile. Ramanujan's legacy in the "Notebooks" promises not only to enrich pure mathematics but also to find application in various fields of mathematical physics. Rodney J. Baxter of the Australian National University, for example, acknowledges that Ramanujan's findings helped him to solve such problems in statistical mechanics as the so-called hard-hexagon model, which considers the behavior of a system of interacting particles laid out on a honeycomblike grid. Similarly, Carlos J. Moreno of the City University of New York and Freeman J. Dyson of the Institute for Advanced Study have pointed out that Ramanujan's work is beginning to be applied by physicists in superstring theory.

Ramanujan's stature as a mathematician is all the more astonishing when one considers his limited formal education. He was born on December 22, 1887, into a somewhat impoverished family of the Brahmin caste in the town of Erode in southern India and grew up in Kumbakonam, where his father was an accountant to a clothier. His mathematical precocity was recognized early, and at the age of seven he was given a scholarship to the Kumbakonam Town High School. He is said to have recited mathematical formulas to his schoolmates—including the value of pi to many places.

When he was 12, Ramanujan mastered the contents of S. L. Loney's rather comprehensive *Plane Trigonometry,* including its discussion of the sum and products of infinite sequences, which later were to figure prominently in his work. (An infinite sequence is an unending string of terms, often generated by a simple formula. In this context the interesting sequences are those whose terms can be added or multiplied to yield an identifiable, finite value. If the terms are added, the resulting expression is called a series; if they are multiplied, it is called a product.) Three years later he borrowed the *Synopsis of Elementary Results in Pure Mathematics,* a listing of some 6,000 theorems (most of them given without proof) compiled by G. S. Carr, a tutor at the University of Cambridge. Those two books were the basis of Ramanujan's mathematical training.

In 1903 Ramanujan was admitted to a local government college. Yet total absorption in his own mathematical diversions at the expense of everything else caused him to fail his examinations, a pattern repeated four years later at another college in Madras. Ramanujan did set his avocation aside—if only temporarily—to look for a job after his marriage in 1909. Fortunately in 1910 R. Ramachandra Rao, a well-to-do patron of mathematics, gave him a monthly stipend largely on the strength of favorable recommendations from various sympathetic Indian mathematicians and the findings he already had jotted down in the "Notebooks."

In 1912, wanting more conventional work, he took a clerical position in the Madras Port Trust, where the chairman was a British engineer, Sir Francis Spring, and the manager was V. Ramaswami Aiyar, the founder of the Indian Mathematical Society. They encouraged Ramanujan to communicate his results to three prominent British mathematicians. Two apparently did not respond; the one who did was G. H. Hardy of Cambridge, now regarded as the foremost British mathematician of the period.

Hardy, accustomed to receiving crank mail, was inclined to disregard Ramanujan's letter at first glance the day it arrived, January 16, 1913. But after dinner that night Hardy and a close colleague, John E. Littlewood, sat down to puzzle through a list of 120 formulas and theorems Ramanujan had appended to his letter. Some hours later they had reached a verdict: they were seeing the work of a genius and not a crackpot. (According to his own "pure-talent scale" of mathematicians, Hardy was later to rate Ramanujan a 100, Littlewood a 30 and himself a 25. The German mathematician David Hilbert, the most influential figure of the time, merited only an 80.) Hardy described the revelation and its consequences as the one romantic incident in his life. He wrote that some of Ramanujan's formulas defeated him completely, and yet "they must be true, because if they were not true, no one would have had the imagination to invent them."

Hardy immediately invited Ramanujan to come to Cambridge. In spite of his mother's strong objections as well as his own reservations, Ramanujan set out for England in March of 1914. During the next five years Hardy and Ramanujan worked together at Trinity College. The blend of Hardy's technical expertise and Ramanujan's raw brilliance produced an unequaled collaboration. They published a series of seminal papers on the properties of various arithmetic functions, laying the groundwork for the answer to such questions as: How many prime divisors is a given number likely to have? How many ways can one express a number as a sum of smaller positive integers?

In 1917 Ramanujan was made a Fellow of the Royal Society of London and a Fellow of Trinity College—the first Indian to be awarded either honor. Yet as his prominence grew his health deteriorated sharply, a decline perhaps accelerated by the difficulty of maintaining a strict vegetarian diet in war-rationed England. Although Ramanujan was in and out of sanatoriums, he continued to pour forth new results. In 1919, when peace made travel abroad safe again, Ramanujan returned to India. Already an icon for young Indian intellectuals, the 32-year-old Ramanujan died on April 26, 1920, of what was then diagnosed as tuberculosis but now is thought to have been a severe vitamin deficiency. True to mathematics until the end, Ramanujan did not slow down during his last, pain-racked months, producing the re-

SRINIVASA RAMANUJAN, born in 1887 in India, managed in spite of limited formal education to reconstruct almost single-handedly much of the edifice of number theory and to go on to derive original theorems and formulas. Like many illustrious mathematicians before him, Ramanujan was fascinated by pi: the ratio of any circle's circumference to its diameter. Based on his investigation of modular equations (*see box on page 114*), he formulated exact expressions for pi and derived from them approximate values. As a result of the work of various investigators (including the authors), Ramanujan's methods are now better understood and have been implemented as algorithms.

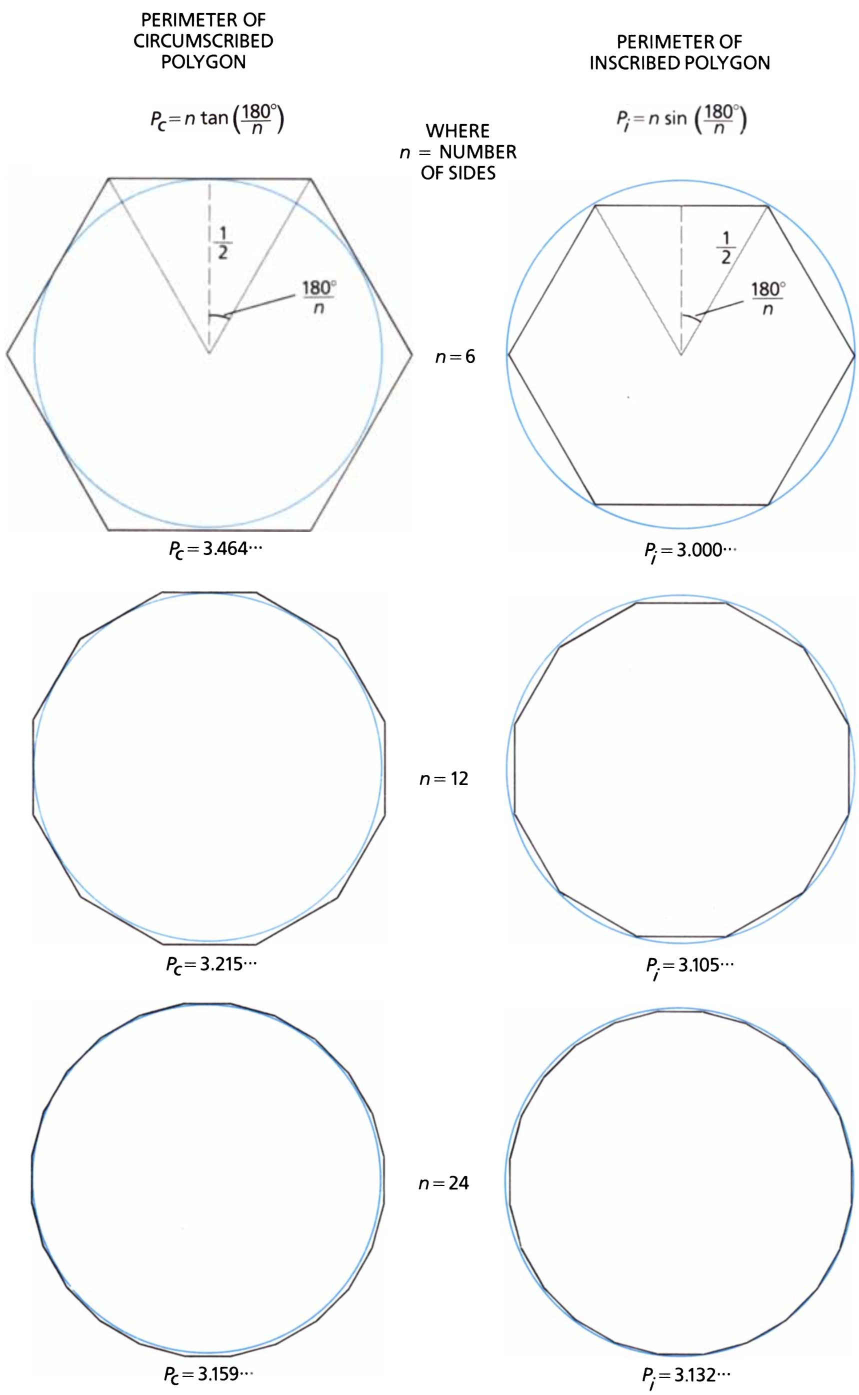
markable work recorded in his so-called "Lost Notebook."

Ramanujan's work on pi grew in large part out of his investigation of modular equations, perhaps the most thoroughly treated subject in the "Notebooks." Roughly speaking, a modular equation is an algebraic relation between a function expressed in terms of a variable x—in mathematical notation, $f(x)$—and the same function expressed in terms of x raised to an integral power, for example $f(x^2)$, $f(x^3)$ or $f(x^4)$. The "order" of the modular equation is given by the integral power. The simplest modular equation is the second-order one: $f(x) = 2\sqrt{f(x^2)}/[1 + f(x^2)]$. Of course, not every function will satisfy a modular equation, but there is a class of functions, called modular functions, that do. These functions have various surprising symmetries that give them a special place in mathematics.

Ramanujan was unparalleled in his ability to come up with solutions to modular equations that also satisfy other conditions. Such solutions are called singular values. It turns out that solving for singular values in certain cases yields numbers whose natural logarithms coincide with pi (times a constant) to a surprising number of places [*see box on page 114*]. Applying this general approach with extraordinary virtuosity, Ramanujan produced many remarkable infinite series as well as single-term approximations for pi. Some of them are given in Ramanujan's one formal paper on the subject, *Modular Equations and Approximations to π*, published in 1914.

Ramanujan's attempts to approximate pi are part of a venerable tradition. The earliest Indo-European civilizations were aware that the area of a circle is proportional to the square of its radius and that the circumference of a circle is directly proportional to its diameter. Less clear, however, is when it was first realized that the ratio of any circle's circumference to its diameter and the ratio of any circle's area to the square of its radius are in fact the same constant, which today is designated by the symbol π. (The symbol, which gives the constant its name, is a latecomer in the history of mathematics, having been introduced in 1706 by the English mathematical writer William Jones and popularized by the Swiss mathematician Leonhard Euler in the 18th century.)

Archimedes of Syracuse, the greatest mathematician of antiquity, rigorously established the equivalence of the two ratios in his treatise *Measurement of a Circle*. He also calculated a value for pi based on mathematical principles rather than on direct measurement of a circle's circumference, area and diameter. What Archimedes did was to inscribe and circumscribe regular polygons (polygons whose sides are all the same length) on a circle assumed to have a diameter of one unit and to consider



ARCHIMEDES' METHOD for estimating pi relied on inscribed and circumscribed regular polygons (polygons with sides of equal length) on a circle having a diameter of one unit (or a radius of half a unit). The perimeters of the inscribed and circumscribed polygons served respectively as lower and upper bounds for the value of pi. The sine and tangent functions can be used to calculate the polygons' perimeters, as is shown here, but Archimedes had to develop equivalent relations based on geometric constructions. Using 96-sided polygons, he determined that pi is greater than $3^{10}/_{71}$ and less than $3^{1}/_{7}$.

the polygons' respective perimeters as lower and upper bounds for possible values of the circumference of the circle, which is numerically equal to pi [*see illustration on opposite page*].

This method of approaching a value for pi was not novel: inscribing polygons of ever more sides in a circle had been proposed earlier by Antiphon, and Antiphon's contemporary, Bryson of Heraclea, had added circumscribed polygons to the procedure. What was novel was Archimedes' correct determination of the effect of doubling the number of sides on both the circumscribed and the inscribed polygons. He thereby developed a procedure that, when repeated enough times, enables one in principle to calculate pi to any number of digits. (It should be pointed out that the perimeter of a regular polygon can be readily calculated by means of simple trigonometric functions: the sine, cosine and tangent functions. But in Archimedes' time, the third century B.C., such functions were only partly understood. Archimedes therefore had to rely mainly on geometric constructions, which made the calculations considerably more demanding than they might appear today.)

Archimedes began with inscribed and circumscribed hexagons, which yield the inequality $3 < \pi < 2\sqrt{3}$. By doubling the number of sides four times, to 96, he narrowed the range of pi to between $3\frac{10}{71}$ and $3\frac{1}{7}$, obtaining the estimate $\pi \approx 3.14$. There is some evidence that the extant text of *Measurement of a Circle* is only a fragment of a larger work in which Archimedes described how, starting with decagons and doubling them six times, he got a five-digit estimate: $\pi \approx 3.1416$.

Archimedes' method is conceptually simple, but in the absence of a ready way to calculate trigonometric functions it requires the extraction of roots, which is rather time-consuming when done by hand. Moreover, the estimates converge slowly to pi: their error decreases by about a factor of four per iteration. Nevertheless, all European attempts to calculate pi before the mid-17th century relied in one way or another on the method. The 16th-century Dutch mathematician Ludolph van Ceulen dedicated much of his career to a computation of pi. Near the end of his life he obtained a 32-digit estimate by calculating the perimeter of inscribed and circumscribed polygons having 2^{62} (some 10^{18}) sides. His value for pi, called the Ludolphian number in parts of Europe, is said to have served as his epitaph.

The development of calculus, largely by Isaac Newton and Gottfried Wilhelm Leibniz, made it possible to calculate pi much more expeditiously. Calculus provides efficient techniques for computing a function's derivative (the rate of change in the function's value as its variables change) and its integral (the sum of the function's values over a range of variables). Applying the techniques, one can demonstrate that inverse trigonometric functions are given by integrals of quadratic functions that describe the curve of a circle. (The inverse of a trigonometric function gives the angle that corresponds to a particular value of the function. For example, the inverse tangent of 1 is 45 degrees or, equivalently, $\pi/4$ radians.)

(The underlying connection between trigonometric functions and algebraic expressions can be appreciated by considering a circle that has a radius of one unit and its center at the origin of a Cartesian *x-y* plane. The equation for the circle—whose area is numerically equal to pi—is $x^2 + y^2 = 1$, which is a restatement of the Pythagorean theorem for a right triangle with a hypotenuse equal to 1. Moreover, the sine and cosine of the angle between the positive *x* axis and any point on the circle are equal respectively to the point's coordinates, *y* and *x;* the angle's tangent is simply *y/x*.)

Of more importance for the purposes of calculating pi, however, is the fact that an inverse trigonometric function can be "expanded" as a series, the terms of which are computable from the derivatives of the function. Newton himself calculated pi to 15 places by adding the first few terms of a series that can be derived

WALLIS' PRODUCT (1665)

$$\frac{\pi}{2} = \frac{2\times2}{1\times3}\times\frac{4\times4}{3\times5}\times\frac{6\times6}{5\times7}\times\frac{8\times8}{7\times9}\times\cdots = \prod_{n=1}^{\infty}\frac{4n^2}{4n^2-1}$$

GREGORY'S SERIES (1671)

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}$$

MACHIN'S FORMULA (1706)

$$\frac{\pi}{4} = 4\arctan(1/5) - \arctan(1/239), \qquad \text{where } \arctan X = X - \frac{X^3}{3} + \frac{X^5}{5} - \frac{X^7}{7} + \cdots = \sum_{n=0}^{\infty}(-1)^n\frac{X^{(2n+1)}}{2n+1}$$

RAMANUJAN (1914)

$$\frac{1}{\pi} = \frac{\sqrt{8}}{9{,}801}\sum_{n=0}^{\infty}\frac{(4n)!\,[1{,}103 + 26{,}390n]}{(n!)^4\,396^{4n}}, \qquad \text{where } n! = n \times (n-1) \times (n-2) \times \cdots \times 1 \text{ and } 0! = 1$$

BORWEIN AND BORWEIN (1987)

$$\frac{1}{\pi} = 12\sum_{n=0}^{\infty}\frac{(-1)^n(6n)!\,[212{,}175{,}710{,}912\sqrt{61} + 1{,}657{,}145{,}277{,}365 + n(13{,}773{,}980{,}892{,}672\sqrt{61} + 107{,}578{,}229{,}802{,}750)]}{(n)!^3(3n)!\,[5{,}280(236{,}674 + 30{,}303\sqrt{61})]^{(3n+3/2)}}$$

TERMS OF MATHEMATICAL SEQUENCES can be summed or multiplied to yield values for pi (divided by a constant) or its reciprocal. The first two sequences, discovered respectively by the mathematicians John Wallis and James Gregory, are probably among the best-known, but they are practically useless for computational purposes. Not even 100 years of computing on a supercomputer programmed to add or multiply the terms of either sequence would yield 100 digits of pi. The formula discovered by John Machin made the calculation of pi feasible, since calculus allows the inverse tangent (arc tangent) of a number, *x*, to be expressed in terms of a sequence whose sum converges more rapidly to the value of the arc tangent the smaller *x* is. Virtually all calculations for pi from the beginning of the 18th century until the early 1970's have relied on variations of Machin's formula. The sum of Ramanujan's sequence converges to the true value of $1/\pi$ much faster: each successive term in the sequence adds roughly eight more correct digits. The last sequence, formulated by the authors, adds about 25 digits per term; the first term (for which *n* is 0) yields a number that agrees with pi to 24 digits.

as an expression for the inverse of the sine function. He later confessed to a colleague: "I am ashamed to tell you to how many figures I carried these calculations, having no other business at the time."

In 1674 Leibniz derived the formula $1 - 1/3 + 1/5 - 1/7 \ldots = \pi/4$, which is the inverse tangent of 1. (The general inverse-tangent series was originally discovered in 1671 by the Scottish mathematician James Gregory. Indeed, similar expressions appear to have been developed independently several centuries earlier in India.) The error of the approximation, defined as the difference between the sum of n terms and the exact value of $\pi/4$, is roughly equal to the $n+1$th term in the series. Since the denominator of each successive term increases by only 2, one must add approximately 50 terms to get two-digit accuracy, 500 terms for three-digit accuracy and so on. Summing the terms of the series to calculate a value for pi more than a few digits long is clearly prohibitive.

An observation made by John Machin, however, made it practicable to calculate pi by means of a series expansion for the inverse-tangent function. He noted that pi divided by 4 is equal to 4 times the inverse tangent of 1/5 minus the inverse tangent of 1/239. Because the inverse-tangent series for a given value converges more quickly the smaller the value is, Machin's formula greatly simplified the calculation. Coupling his formula with the series expansion for the inverse tangent, Machin computed 100 digits of pi in 1706. Indeed, his technique proved to be so powerful that all extended calculations of pi from the beginning of the 18th century until recently relied on variants of the method.

Two 19th-century calculations deserve special mention. In 1844 Johann Dase computed 205 digits of pi in a matter of months by calculating the values of three inverse tangents in a Machin-like formula. Dase was a calculating prodigy who could multiply 100-digit numbers entirely in his head—a feat that took him roughly eight hours. (He was perhaps the closest precursor of the modern supercomputer, at least in terms of memory capacity.) In 1853 William Shanks outdid Dase by publishing his computation of pi to 607 places, although the digits that followed the 527th place were wrong. Shank's task took years and was a rather routine, albeit laborious, application of Machin's formula. (In what must itself be some kind of record, 92 years passed before Shank's error was detected, in a comparison between his value and a 530-place approximation produced by D. F. Ferguson with the aid of a mechanical calculator.)

The advent of the digital computer saw a renewal of efforts to calculate ever more digits of pi, since the machine was ideally suited for lengthy, repetitive "number crunching." ENIAC, one of the first digital computers, was applied to the task in June, 1949, by John von Neumann and his colleagues. ENIAC produced 2,037 digits in 70 hours. In 1957 G. E. Felton attempted to compute 10,000 digits of pi, but owing to a machine error only the first 7,480 digits were correct. The 10,000-digit goal was reached by F. Genuys the following year on an IBM 704 computer. In 1961 Daniel Shanks and John W. Wrench, Jr., calculated 100,000 digits of pi in less than nine hours on an IBM 7090. The million-digit mark was passed in 1973 by Jean Guilloud and M. Bouyer, a feat that took just under a day of computation on a CDC 7600. (The computations done by Shanks and Wrench and by Guilloud and Bouyer were in fact carried out twice using different inverse-tangent identities for pi. Given the history of both human and machine error in these calculations, it is only after such verification that modern "digit hunters" consider a record officially set.)

Although an increase in the speed of computers was a major reason ever more accurate calculations for pi could be performed, it soon became clear that there were inescapable limits. Doubling the number of digits lengthens computing time by at least a factor of four, if one applies the traditional methods of performing arithmetic in computers. Hence even allowing for a hundredfold increase in computational speed, Guilloud and Bouyer's program would have required at least a quarter century to produce a billion-digit value for pi. From the perspective of the early 1970's such a computation did not seem realistically practicable.

Yet the task is now feasible, thanks

MODULAR FUNCTIONS AND APPROXIMATIONS TO PI

A modular function is a function, $\lambda(q)$, that can be related through an algebraic expression called a modular equation to the same function expressed in terms of the same variable, q, raised to an integral power: $\lambda(q^p)$. The integral power, p, determines the "order" of the modular equation. An example of a modular function is

$$\lambda(q) = 16q \prod_{n=1}^{\infty} \left(\frac{1+q^{2n}}{1+q^{2n-1}}\right)^8 .$$

Its associated seventh-order modular equation, which relates $\lambda(q)$ to $\lambda(q^7)$, is given by

$$\sqrt[8]{\lambda(q)\lambda(q^7)} + \sqrt[8]{[1-\lambda(q)][1-\lambda(q^7)]} = 1 .$$

Singular values are solutions of modular equations that must also satisfy additional conditions. One class of singular values corresponds to computing a sequence of values, k_p, where

$$k_p = \sqrt{\lambda(e^{-\pi\sqrt{p}})}$$

and p takes integer values. These values have the curious property that the logarithmic expression

$$\frac{-2}{\sqrt{p}} \log\left(\frac{k_p}{4}\right)$$

coincides with many of the first digits of pi. The number of digits the expression has in common with pi increases with larger values of p.

Ramanujan was unparalleled in his ability to calculate these singular values. One of his most famous is the value when p equals 210, which was included in his original letter to G. H. Hardy. It is

$$k_{210} = (\sqrt{2}-1)^2(2-\sqrt{3})(\sqrt{7}-\sqrt{6})^2(8-3\sqrt{7})(\sqrt{10}-3)^2(\sqrt{15}-\sqrt{14})(4-\sqrt{15})^2(6-\sqrt{35}) .$$

This number, when plugged into the logarithmic expression, agrees with pi through the first 20 decimal places. In comparison, $k_{2^{40}}$ yields a number that agrees with pi through more than one million digits.

Applying this general approach, Ramanujan constructed a number of remarkable series for pi, including the one shown in the illustration on the preceding page. The general approach also underlies the two-step, iterative algorithms in the top illustration on the opposite page. In each iteration the first step (calculating y_n) corresponds to computing one of a sequence of singular values by solving a modular equation of the appropriate order; the second step (calculating α_n) is tantamount to taking the logarithm of the singular value.

not only to faster computers but also to new, efficient methods for multiplying large numbers in computers. A third development was also crucial: the advent of iterative algorithms that quickly converge to pi. (An iterative algorithm can be expressed as a computer program that repeatedly performs the same arithmetic operations, taking the output of one cycle as the input for the next.) These algorithms, some of which we constructed, were in many respects anticipated by Ramanujan, although he knew nothing of computer programming. Indeed, computers not only have made it possible to apply Ramanujan's work but also have helped to unravel it. Sophisticated algebraic-manipulation software has allowed further exploration of the road Ramanujan traveled alone and unaided 75 years ago.

One of the interesting lessons of theoretical computer science is that many familiar algorithms, such as the way children are taught to multiply in grade school, are far from optimal. Computer scientists gauge the efficiency of an algorithm by determining its bit complexity: the number of times individual digits are added or multiplied in carrying out an algorithm. By this measure, adding two n-digit numbers in the normal way has a bit complexity that increases in step with n; multiplying two n-digit numbers in the normal way has a bit complexity that increases as n^2. By traditional methods, multiplication is much "harder" than addition in that it is much more time-consuming.

Yet, as was shown in 1971 by A. Schönhage and V. Strassen, the multiplication of two numbers can in theory have a bit complexity only a little greater than addition. One way to achieve this potential reduction in bit complexity is to implement so-called fast Fourier transforms (FFT's). FFT-based multiplication of two large numbers allows the intermediary computations among individual digits to be carefully orchestrated so that redundancy is avoided. Because division and root extraction can be reduced to a sequence of multiplications, they too can have a bit complexity just slightly greater than that of addition. The result is a tremendous saving in bit complexity and hence in computation time. For this reason all recent efforts to calculate pi rely on some variation of the FFT technique for multiplication.

Yet for hundreds of millions of dig-

(a) Let $y_0 = \frac{1}{\sqrt{2}}$ $\alpha_0 = \frac{1}{2}$

and

$$y_{n+1} = \frac{1-\sqrt{1-y_n^2}}{1+\sqrt{1-y_n^2}} \qquad \alpha_{n+1} = [(1+y_{n+1})^2\alpha_n] - 2^{n+1}y_{n+1}$$

(b) Let $y_0 = \sqrt{2}-1$ $\alpha_0 = 6-4\sqrt{2}$

and

$$y_{n+1} = \frac{1-\sqrt[4]{1-y_n^4}}{1+\sqrt[4]{1-y_n^4}}$$

$$\alpha_{n+1} = [(1+y_{n+1})^4\alpha_n] - 2^{2n+3}y_{n+1}(1+y_{n+1}+y_{n+1}^2)$$

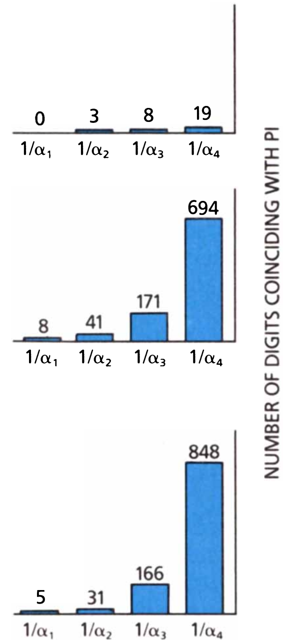
(c) Let $s_0 = 5(\sqrt{5}-2)$ $\alpha_0 = \frac{1}{2}$

and

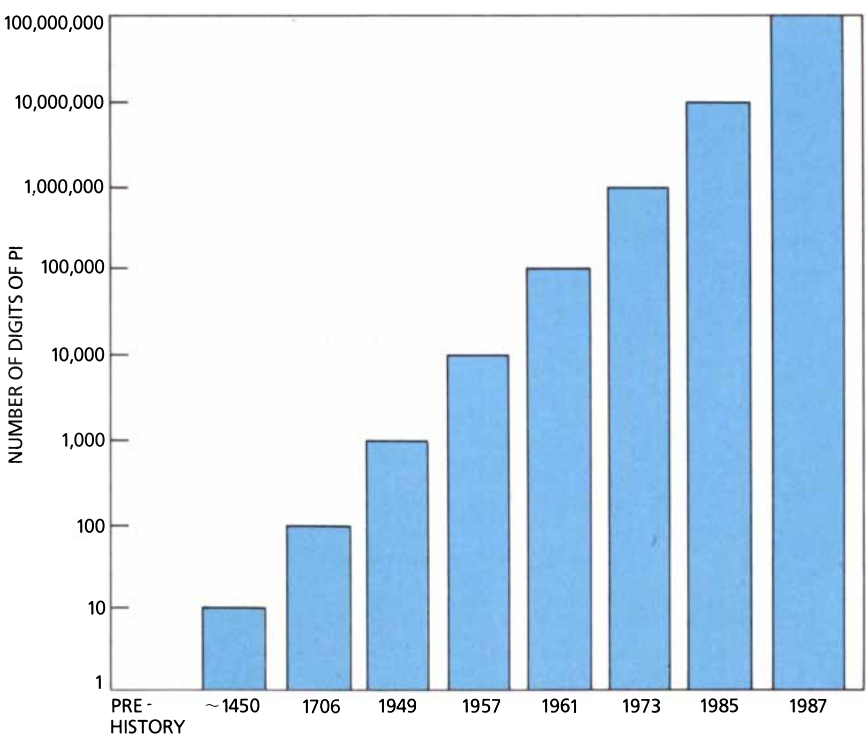
$$s_{n+1} = \frac{25}{s_n(Z+X/Z+1)^2}, \text{ where } X = \frac{5}{s_n}-1, \quad Y = (X-1)^2+7$$

$$\text{and } Z = \sqrt[5]{\frac{X(Y+\sqrt{Y^2-4X^3})}{2}}$$

$$\alpha_{n+1} = s_n^2\alpha_n - 5^n\left[\frac{s_n^2-5}{2} + \sqrt{s_n(s_n^2-2s_n+5)}\right]$$



ITERATIVE ALGORITHMS that yield extremely accurate values of pi were developed by the authors. (An iterative algorithm is a sequence of operations repeated in such a way that the ouput of one cycle is taken as the input for the next.) Algorithm *a* converges to $1/\pi$ quadratically: the number of correct digits given by α_n more than doubles each time n is increased by 1. Algorithm *b* converges quartically and algorithm *c* converges quintically, so that the number of coinciding digits given by each iteration increases respectively by more than a factor of four and by more than a factor of five. Algorithm *b* is possibly the most efficient known algorithm for calculating pi; it was run on supercomputers in the last three record-setting calculations. As the authors worked on the algorithms it became clear to them that Ramanujan had pursued similar methods in coming up with his approximations for pi. In fact, the computation of s_n in algorithm *c* rests on a remarkable fifth-order modular equation discovered by Ramanujan.



NUMBER OF KNOWN DIGITS of pi has increased by two orders of magnitude (factors of 10) in the past decade as a result of the development of iterative algorithms that can be run on supercomputers equipped with new, efficient methods of multiplication.

its of pi to be calculated practically a beautiful formula known a century and a half earlier to Carl Friedrich Gauss had to be rediscovered. In the mid-1970's Richard P. Brent and Eugene Salamin independently noted that the formula produced an algorithm for pi that converged quadratically, that is, the number of digits doubled with each iteration. Between 1983 and the present Yasumasa Kanada and his colleagues at the University of Tokyo have employed this algorithm to set several world records for the number of digits of pi.

We wondered what underlies the remarkably fast convergence to pi of the Gauss-Brent-Salamin algorithm, and in studying it we developed general techniques for the construction of similar algorithms that rapidly converge to pi as well as to other quantities. Building on a theory outlined by the German mathematician Karl Gustav Jacob Jacobi in 1829, we realized we could in principle arrive at a value for pi by evaluating integrals of a class called elliptic integrals, which can serve to calculate the perimeter of an ellipse. (A circle, the geometric setting of previous efforts to approximate pi, is simply an ellipse with axes of equal length.)

Elliptic integrals cannot generally be evaluated as integrals, but they can be easily approximated through iterative procedures that rely on modular equations. We found that the Gauss-Brent-Salamin algorithm is actually a specific case of our more general technique relying on a second-order modular equation. Quicker convergence to the value of the integral, and thus a faster algorithm for pi, is possible if higher-order modular equations are used, and so we have also constructed various algorithms based on modular equations of third, fourth and higher orders.

In January, 1986, David H. Bailey of the National Aeronautics and Space Administration's Ames Research Center produced 29,360,000 decimal places of pi by iterating one of our algorithms 12 times on a Cray-2 supercomputer. Because the algorithm is based on a fourth-order modular equation, it converges on pi quartically, more than quadrupling the number of digits with each iteration. A year later Kanada and his colleagues carried out one more iteration to attain 134,217,000 places on an NEC SX-2 supercomputer and thereby verified a similar computation they had done earlier using the Gauss-Brent-Salamin algorithm. (Iterating our algorithm twice more—a feat entirely feasible if one could somehow monopolize a supercomputer for a few weeks—would yield more than two billion digits of pi.)

HOW TO GET TWO BILLION DIGITS OF PI WITH A CALCULATOR*

Let

$y_0 = \sqrt{2} - 1$	$\alpha_0 = 6 - 4\sqrt{2}$
$y_1 = [1 - \sqrt[4]{1 - y_0^4}]/[1 + \sqrt[4]{1 - y_0^4}]$	$\alpha_1 = (1 + y_1)^4 \alpha_0 - 2^3 y_1(1 + y_1 + y_1^2)$
$y_2 = [1 - \sqrt[4]{1 - y_1^4}]/[1 + \sqrt[4]{1 - y_1^4}]$	$\alpha_2 = (1 + y_2)^4 \alpha_1 - 2^5 y_2(1 + y_2 + y_2^2)$
$y_3 = [1 - \sqrt[4]{1 - y_2^4}]/[1 + \sqrt[4]{1 - y_2^4}]$	$\alpha_3 = (1 + y_3)^4 \alpha_2 - 2^7 y_3(1 + y_3 + y_3^2)$
$y_4 = [1 - \sqrt[4]{1 - y_3^4}]/[1 + \sqrt[4]{1 - y_3^4}]$	$\alpha_4 = (1 + y_4)^4 \alpha_3 - 2^9 y_4(1 + y_4 + y_4^2)$
$y_5 = [1 - \sqrt[4]{1 - y_4^4}]/[1 + \sqrt[4]{1 - y_4^4}]$	$\alpha_5 = (1 + y_5)^4 \alpha_4 - 2^{11} y_5(1 + y_5 + y_5^2)$
$y_6 = [1 - \sqrt[4]{1 - y_5^4}]/[1 + \sqrt[4]{1 - y_5^4}]$	$\alpha_6 = (1 + y_6)^4 \alpha_5 - 2^{13} y_6(1 + y_6 + y_6^2)$
$y_7 = [1 - \sqrt[4]{1 - y_6^4}]/[1 + \sqrt[4]{1 - y_6^4}]$	$\alpha_7 = (1 + y_7)^4 \alpha_6 - 2^{15} y_7(1 + y_7 + y_7^2)$
$y_8 = [1 - \sqrt[4]{1 - y_7^4}]/[1 + \sqrt[4]{1 - y_7^4}]$	$\alpha_8 = (1 + y_8)^4 \alpha_7 - 2^{17} y_8(1 + y_8 + y_8^2)$
$y_9 = [1 - \sqrt[4]{1 - y_8^4}]/[1 + \sqrt[4]{1 - y_8^4}]$	$\alpha_9 = (1 + y_9)^4 \alpha_8 - 2^{19} y_9(1 + y_9 + y_9^2)$
$y_{10} = [1 - \sqrt[4]{1 - y_9^4}]/[1 + \sqrt[4]{1 - y_9^4}]$	$\alpha_{10} = (1 + y_{10})^4 \alpha_9 - 2^{21} y_{10}(1 + y_{10} + y_{10}^2)$
$y_{11} = [1 - \sqrt[4]{1 - y_{10}^4}]/[1 + \sqrt[4]{1 - y_{10}^4}]$	$\alpha_{11} = (1 + y_{11})^4 \alpha_{10} - 2^{23} y_{11}(1 + y_{11} + y_{11}^2)$
$y_{12} = [1 - \sqrt[4]{1 - y_{11}^4}]/[1 + \sqrt[4]{1 - y_{11}^4}]$	$\alpha_{12} = (1 + y_{12})^4 \alpha_{11} - 2^{25} y_{12}(1 + y_{12} + y_{12}^2)$
$y_{13} = [1 - \sqrt[4]{1 - y_{12}^4}]/[1 + \sqrt[4]{1 - y_{12}^4}]$	$\alpha_{13} = (1 + y_{13})^4 \alpha_{12} - 2^{27} y_{13}(1 + y_{13} + y_{13}^2)$
$y_{14} = [1 - \sqrt[4]{1 - y_{13}^4}]/[1 + \sqrt[4]{1 - y_{13}^4}]$	$\alpha_{14} = (1 + y_{14})^4 \alpha_{13} - 2^{29} y_{14}(1 + y_{14} + y_{14}^2)$
$y_{15} = [1 - \sqrt[4]{1 - y_{14}^4}]/[1 + \sqrt[4]{1 - y_{14}^4}]$	$\alpha_{15} = (1 + y_{15})^4 \alpha_{14} - 2^{31} y_{15}(1 + y_{15} + y_{15}^2)$

$1/\alpha_{15}$ agress with π for more than two billion decimal digits

*Of course, the calculator needs to have a two-billion-digit display; on a pocket calculator the computation would not be very interesting after the second iteration.

EXPLICIT INSTRUCTIONS for executing algorithm *b* in the top illustration on the preceding page makes it possible in principle to compute the first two billion digits of pi in a matter of minutes. All one needs is a calculator that has two memory registers and the usual capacity to add, subtract, multiply, divide and extract roots. Unfortunately most calculators come with only an eight-digit display, which makes the computation moot.

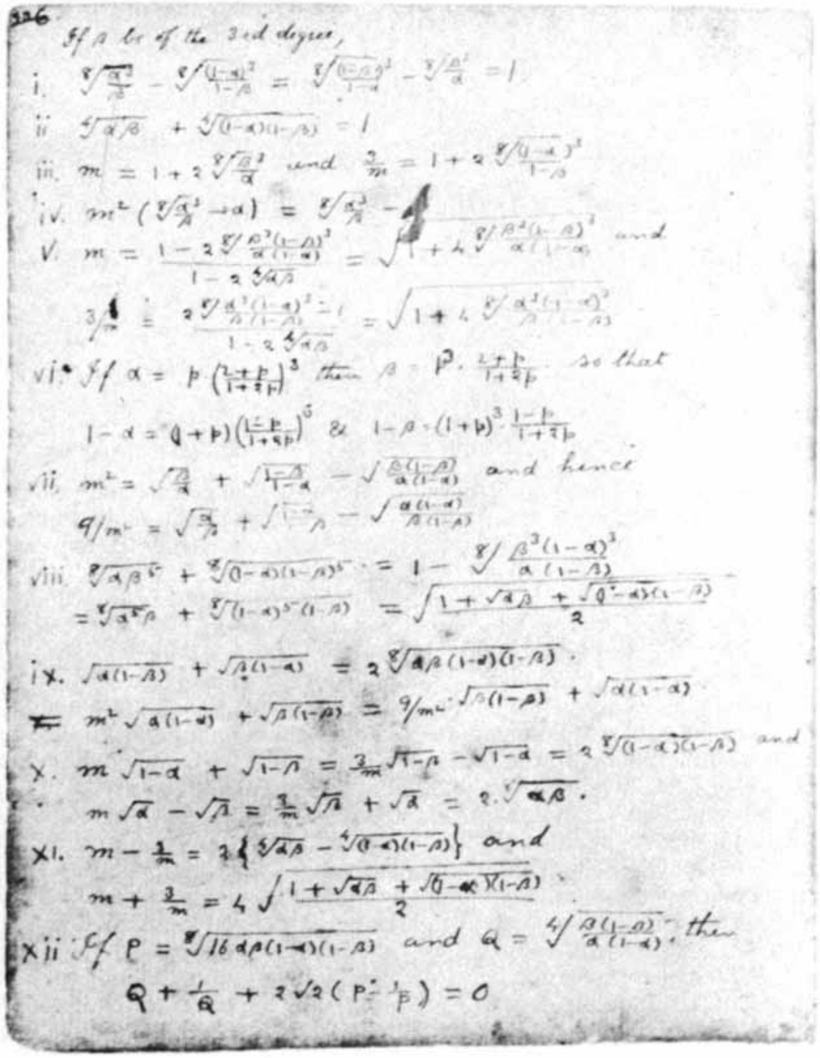
Iterative methods are best suited for calculating pi on a computer, and so it is not surprising that Ramanujan never bothered to pursue them. Yet the basic ingredients of the iterative algorithms for pi—modular equations in particular—are to be found in Ramanujan's work. Parts of his original derivation of infinite series and approximations for pi more than three-quarters of a century ago must have paralleled our own efforts to come up with algorithms for pi. Indeed, the formulas he lists in his paper on pi and in the "Notebooks" helped us greatly in the construction of some of our algorithms. For example, although we were able to prove that an 11th-order algorithm exists and knew its general formulation, it was not until we stumbled on Ramanujan's modular equations of the same order that we discovered its unexpectedly simple form.

Conversely, we were also able to derive all Ramanujan's series from the general formulas we had developed. The derivation of one, which converged to pi faster than any other series we knew at the time, came about with a little help from an unexpected source. We had justified all the quantities in the expression for the series except one: the coefficient 1,103, which appears in the numerator of the expression [*see illustration on page 113*]. We were convinced—as Ramanujan must have been—that 1,103 had to be correct. To prove it we had either to simplify a daunting equation containing variables raised to powers of several thousand or to delve considerably further into somewhat arcane number theory.

By coincidence R. William Gosper, Jr., of Symbolics, Inc., had decided in 1985 to exploit the same series of Ramanujan's for an extended-accuracy value for pi. When he carried out the calculation to more than 17 million digits (a record at the time), there was to his knowledge no proof that the sum of the series actually converged to pi. Of course, he knew that millions of digits of his value coincided with an earlier Gauss-Brent-Salamin calculation done by Kanada. Hence the possibility of error was vanishingly small.

As soon as Gosper had finished his calculation and verified it against Kanada's, however, we had what we needed to prove that 1,103 was the number needed to make the series true to within one part in $10^{10,000,000}$. In much the same way that a pair of integers differing by less than 1 must be equal, his result sufficed to specify the number: it is precisely 1,103. In effect, Gosper's computation became part of our proof. We knew that the series (and its associated algorithm) is so sensitive to slight inaccuracies that if Gosper had used any other value for the coefficient or, for that matter, if the computer had introduced a single-digit error during the calculation, he would have ended up with numerical nonsense instead of a value for pi.

Ramanujan-type algorithms for approximating pi can be shown to be very close to the best possible. If all the operations involved in the execution of the algorithms are totaled (assuming that the best techniques known for addition, multiplication and root extraction are applied), the bit complexity of computing n digits of pi is only marginally greater than that of multiplying two n-digit numbers. But multiplying two n-digit numbers by means of an FFT-based technique is only marginally more complicated than summing two n-digit numbers, which is the simplest of the arithmetic operations possible on a computer.



RAMANUJAN'S "NOTEBOOKS" were personal records in which he jotted down many of his formulas. The page shown contains various third-order modular equations—all in Ramanujan's nonstandard notation. Unfortunately Ramanujan did not bother to include formal proofs for the equations; others have had to compile, edit and prove them. The formulas in the "Notebooks" embody subtle relations among numbers and functions that can be applied in other fields of mathematics or even in theoretical physics.

Mathematics has probably not yet felt the full impact of Ramanujan's genius. There are many other wonderful formulas contained in the "Notebooks" that revolve around integrals, infinite series and continued fractions (a number plus a fraction, whose denominator can be expressed as a number plus a fraction, whose denominator can be expressed as a number plus a fraction, and so on). Unfortunately they are listed with little—if any—indication of the method by which Ramanujan proved them. Littlewood wrote: "If a significant piece of reasoning occurred somewhere, and the total mixture of evidence and intuition gave him certainty, he looked no further."

The herculean task of editing the "Notebooks," initiated 60 years ago by the British analysts G. N. Watson and B. N. Wilson and now being completed by Bruce Berndt, requires providing a proof, a source or an occasional correction for each of many thousands of asserted theorems and identities. A single line in the "Notebooks" can easily elicit many pages of commentary. The task is made all the more difficult by the nonstandard mathematical notation in which the formulas are written. Hence a great deal of Ramanujan's work will not become accessible to the mathematical community until Berndt's project is finished.

Ramanujan's unique capacity for working intuitively with complicated formulas enabled him to plant seeds in a mathematical garden (to borrow a metaphor from Freeman Dyson) that is only now coming into bloom. Along with many other mathematicians, we look forward to seeing which of the seeds will germinate in future years and further beautify the garden.

The Bubonic Plague

A bacterial disease carried by fleas that feed on rats, it has afflicted human beings for more than 1,000 years. The factors responsible for its alternate rise and fall remain a mystery

by Colin McEvedy

In the year 1346 Europe, northern Africa and nearer parts of the Middle East had a total population of approximately 100 million people. In the course of the next few years a fourth of them died, victims of a new and terrifying illness that spread throughout the area, killing most of those unfortunate enough to catch it. The disease put an end to the population rise that had marked the evolution of medieval society: within four years Europe alone suffered a loss of roughly 20 million people. The disease responsible for such grim statistics was the bubonic plague, and this particular outbreak, lasting from 1346 to 1352, was known as the Great Dying or the Great Pestilence. Later it was appropriately referred to as the Black Death, a name that has come down through history.

Although the effects of the Black Death may have been particularly catastrophic, striking as it did after a long period in which the disease had been unknown in the West, this was not the first time the plague had ravaged Europe. Some 800 years earlier, during the reign of the emperor Justinian in the sixth century, there was an epidemic of similar proportions. There were also repeated, if less widespread, epidemics in the two centuries following the plague of Justinian's time, and for four centuries after the Black Death. The disease has undergone a precipitous decline since that time, but it still occurs sporadically in various parts of the world today, including the U.S.

From 70 to 80 percent of those who contracted the plague in the 14th century died from it. Indeed, the symptoms usually presented themselves with a ferocity that presaged death within five days. The name, bubonic plague, derives from one of the early signs of the disease: the appearance of large, painful swellings called buboes in the lymph nodes of the armpit, neck or groin of the victim. Three days after the appearance of the buboes people were characteristically overwhelmed by high fever, became delirious and broke out in black splotches that were the result of hemorrhaging under the skin. As the disease progressed, the buboes continued to grow larger and more painful; often they burst.

The bursting is said to have been particularly agonizing, capable of arousing even the most moribund patients to a state of frenzy. Yet physicians always regarded the bursting as a good sign, if only because it indicated that the patient was still capable of putting up a fight a week or so after the onset of the illness. Of those who were going to die, probably half were already dead by this stage.

In some cases a person's bloodstream was directly infected, which led to septic shock, massive hemorrhaging and rapid death, a form of the disease known as septicemic plague. In other cases plague was transmitted as a type of pneumonia; in pneumonic plague the victims collapsed, spit blood and were almost always dead within a few days.

Strange as it may seem, in view of the frequency of the disease and the toll it exerted on the population, no one at the time had an inkling of its fundamental nature, its ultimate cause or how it was spread. During the period of the Black Death people were inclined to attribute the disease to unfavorable astrological combina-

HORROR OF THE PLAGUE is captured in Pieter Brueghel the Elder's painting "Triumph of Death," where death, in the form of roving skeletons, overwhelms a kingdom of the living. Neither the king with his piles of gold nor the young revelers at their table can escape the relentless army of the dead. Behind the king a corps of skeletons pushes victims into a water-filled mass grave; a barren landscape already robbed of life can be seen in the distance. Apocalyptic visions of this kind were common in the centuries when the plague ravaged Europe and the healthiest could be wiped out in a few days.

tions or malignant atmospheres ("miasmas"), neither of which could be translated into a public-health program of any kind. More paranoid elaborations blamed the disease on deliberate contamination by witches, Moslems (an idea proposed by Christians), Christians (proposed by Moslems) or Jews (proposed by both groups).

It was not until 1894 that the French bacteriologist Alexandre Yersin discovered that bubonic plague is caused by a gram-negative bacterium, *Yersinia pestis,* belonging to a group of bacteria known as rod-shaped bacilli, many of which are pathogenic. Plague bacilli are found at low frequency in many wild rodent populations throughout the world and are transmitted from one rodent to another by fleas. In the case of the bubonic plague the flea often responsible for transmitting the disease is the oriental rat flea, *Xenopsylla cheopis.* When a flea bites an infected rat, it ingests the bacilli, which proceed to replicate within its digestive tract, forming a solid mass that obstructs the flea's gut; the flea is unable to ingest blood and becomes ravenously hungry. In a feeding frenzy it repeatedly bites its animal host, regurgitating plague bacilli into the host's bloodstream every time it does so. These injection sites then act as foci for the spread of bacilli. If the host animal dies, as it is likely to do, the flea moves to the next available live rodent. The disease spreads rapidly in this manner; as the number of live rats decreases, the fleas move to warm-blooded hosts on which they would not normally feed, such as human beings and their domesticated animals, and so an epidemic is launched.

Once the disease enters the human population it can sometimes spread directly from human to human through the inhalation of infected respiratory droplets. The normal mode of spread is by the bite of rat fleas, however; the disease does not persist in the absence of rodents, which are the primary hosts for both the plague bacillus and the rat flea.

The essential requirement for an epidemic (an outbreak in a human population) is a rodent epizootic (an outbreak in an animal population). This is necessary both to initiate and to sustain the disease in human beings. Of course, the two populations must be in close contact for the transmission to be successful, but it is unlikely that this was ever a significant variable in medieval times. In rural as well as urban areas humans lived surrounded by rats.

The Black Death is thought to have migrated along the Silk Road, the trans-Asian route by which Chinese silk was brought to Europe. There are two reasons for believing this was the case. The first is that outbreaks of the plague were recorded in 1346 in Astrakhan and Saray, both caravan stations on the lower Volga River in what is now the U.S.S.R. The second

is that during the years 1347 and 1348 the Arab traveler and scholar Ibn Battuta, returning along the Spice Route from a stay in India, first reported hearing news of the plague when he reached Aleppo in northern Syria, not before. That clue excludes transmission by way of the Indian Ocean and Persian Gulf ports.

Most likely the disease erupted first among marmots, large rodents native to central Asia (they are related to woodchucks but belong to a different species) whose fur was an important article of trade throughout that part of the world. According to this historical reconstruction, trappers coming across dead or dying animals collected their furs, delighted to find such an abundant supply, and sold them to dealers who in turn (without worrying about reports of illness among the fur trappers) sold them to buyers from the West. When the bales of marmot furs sent west along the Silk Road were first opened in Astrakhan and Saray, hungry fleas jumped from the fur, seeking the first available blood meal they could find.

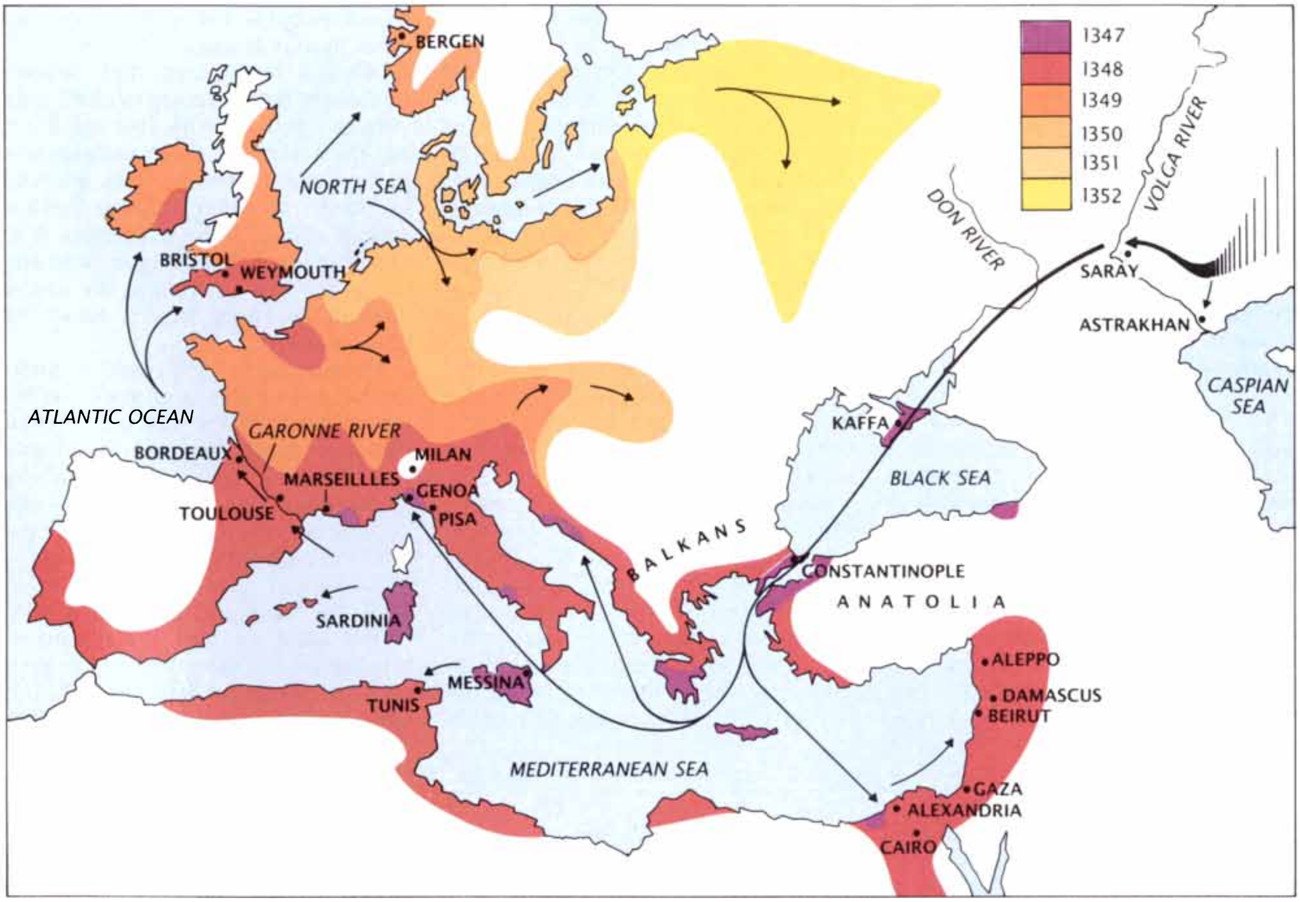
PHYSICIAN'S GARB that was worn during a plague outbreak in Marseilles in 1720. The birdlike costume, made of leather, covered its wearer from head to toe and was believed to provide protection from contagion. The large beak contained sweet-smelling herbs to filter airborne contagion; the wand contained incense that was thought to ward off impurities. Even the eyeholes, which held crystalline lenses, were protective.

From Saray the disease is thought to have traveled down the Don River to Kaffa, a major port on the Black Sea, where a large rat population provided the perfect breeding ground for the plague bacillus. Because many of the rats in Kaffa were living on sailing vessels bound for the ports of Europe, the disease had a ready means of transport to that part of the world.

Indeed, it would be difficult to design a more efficient means of disseminating the plague than a medieval ship. The holds of these ships were generally crawling with rats; when the crew slept, the rats took over, running through the rigging and dropping fleas onto the decks below. The cycle of infection, from flea to rat and rat to flea, would be maintained until the rat population was so reduced by the disease that it could no longer sustain the fleas and the plague bacteria they were carrying. Hungry fleas, seeking any host they could find, would then carry the disease into the human population. It is small wonder that by the end of 1347 plague had broken out in most ports on the route linking Kaffa to Genoa in northern Italy.

The two most important ports along this route were Pera, a suburb of Constantinople, and Messina in Sicily. Both places were stopover points for ships crossing the Mediterranean and became major foci for further dissemination of the plague. The initial impact on the population of Constantinople was graphically described by the emperor Cantacuzenus, who lost a son to the disease in 1347. He recounts how it spread throughout the Greek islands and along the coasts of Anatolia and the Balkans, killing "most of the people." In Messina the first outbreak was recorded in October of 1347, launching an epidemic that quickly spread to include the entire island.

From there in early 1348 the Black Death crossed over to Tunis on the north coast of Africa and then spread by way of Sardinia to Spain. By the time it reached Spain the Black Death had also spread to the heart of Europe, a fact that can be blamed at least in part on the Genoese, who are said to have heartlessly turned away ships from the east carrying their sick countrymen. Not only did such hardheartedness have little effect (the city was as badly hit as any in Europe) but also the diversion of ships to other ports, such as Marseilles and Pisa, hastened the spread of the plague throughout Europe.



BLACK DEATH came from central Asia to Europe via the Silk Road, arriving in Kaffa in about 1347. From there it was carried by ship to the major ports of Europe and northern Africa. Most of Europe was affected before the epidemic finally subsided in 1352. Milan, the largest city to escape the plague, is believed to have done so because it is the farthest Italian city from the sea.

By this time the epidemic was raging throughout the Mediterranean. Ships carrying silk, slaves and fur brought it to Alexandria before the end of 1347; from there it spread south to Cairo, east to Gaza, Beirut and Damascus and finally along the north coast of Africa to Morocco.

By 1348 the Black Death had jumped from the Mediterranean region to the Atlantic coast of Europe. It crossed southwestern France by way of the regional capital, Toulouse, and rapidly passed down the Garonne River to Bordeaux on the west coast. From there it is likely that one of the ships loading claret for the British market brought the Black Death to Great Britain. In 1348 it was first recorded at Weymouth on the south coast of England, and it is believed to have spread to Ireland from Bristol.

From England the plague crossed the North Sea to envelop Scandinavia in its deadly grasp. According to one story, the invasion of Scandinavia can be blamed on a ship that left London in May, 1349, bound for Bergen with a full crew and a cargo of wool. The ship is reported to have been seen some days later, drifting off the coast of Norway. Local people who rowed out to investigate found the crew dead and returned to shore, carrying the wool and—unwittingly—the plague with them. That started a chain reaction as village after village along the Norwegian coast succumbed to the disease.

The following year the Black Death ravaged the populations of Denmark and Germany before entering Poland in 1351 and Russia in 1352. This in effect completed the circle; not only had the disease returned to within a few hundred miles of its entry into Europe on the Volga steppe but also after four long and devastating years mortality rates in western Europe had finally returned to normal.

The society that emerged from the period of the Black Death became quite prosperous; the survivors had inherited the fortunes of their deceased relatives and many were able to move into positions of prominence once closed to them. Their good fortune did not necessarily last for long, however. In 1356 a second outbreak of the plague appeared in Germany and spread rapidly throughout Europe. It exacted a particularly heavy toll among the children born since the end of the Black Death.

Thereafter the plague returned to Europe with mournful regularity; indeed, the continent never seemed free of it for more than a few years at a time. Although the later epidemics never matched the Black Death in terms of overall mortality, they nonetheless continued to have a negative impact on population growth in Europe through the end of the 14th century.

At this point an equilibrium was reached between plague and people, and in the 15th century the population began to recover. In particularly hard-hit regions it took more than a century for numbers to return to their original levels, but by the end of the 16th century populations all over were higher than they had been prior to the onset of the Black Death.

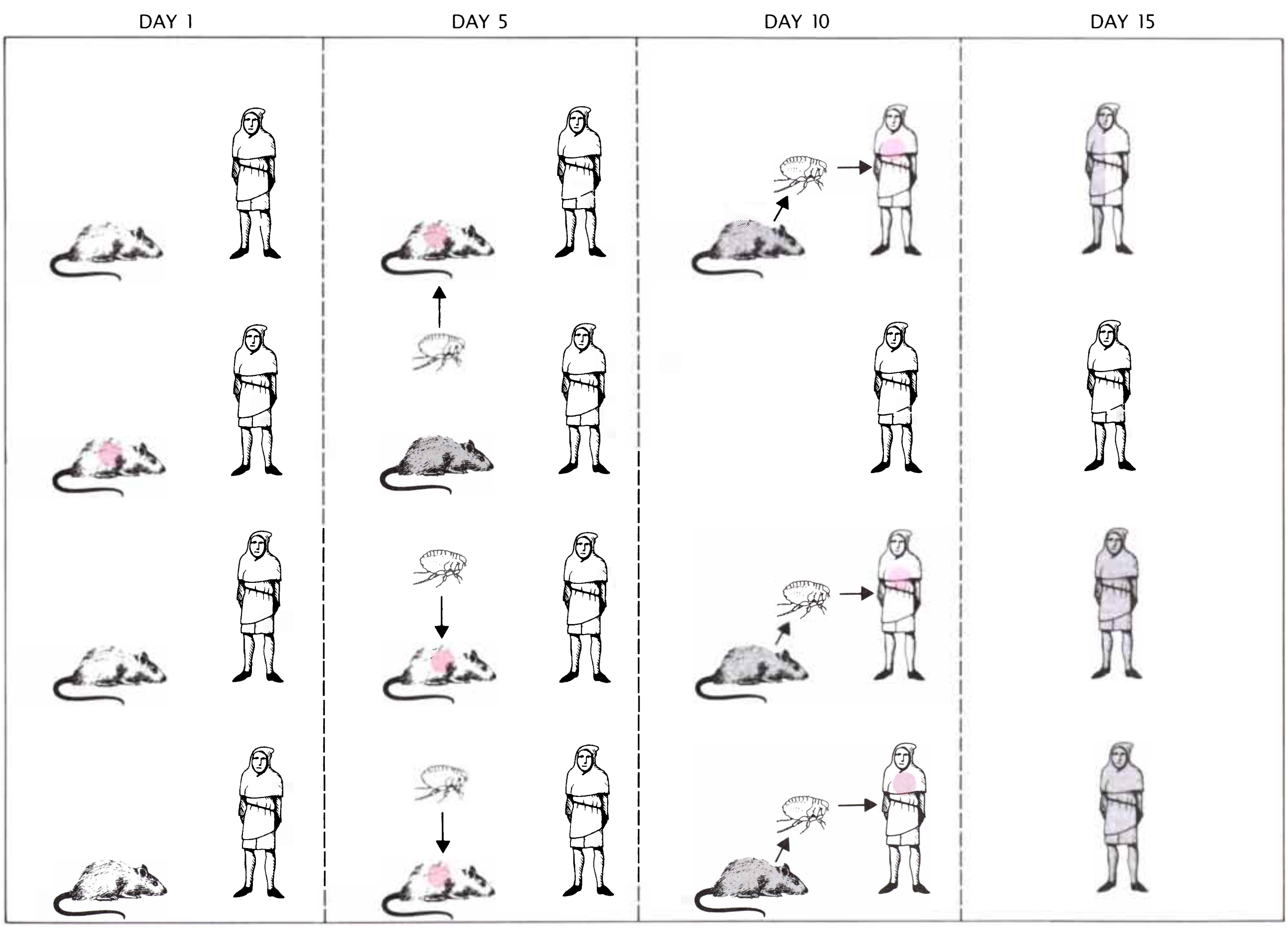
Strangely, when the plague did reappear (which it continued to do, albeit less frequently), it often did so with a ferocity equal to any recorded in previous outbreaks. In the last epidemic in France, from 1720 to 1722, half of the population of the city of Marseilles died, together with 60 percent of the population in neighboring Toulon, 44 percent at Arles and 30 percent at Aix and Avignon. Yet the epidemic did not spread beyond Provence, and the total number of deaths was less than 100,000.

By the 16th century it was widely believed the plague spread as a result of contagion: a toxic factor that could be transferred from the sick to the healthy. Human-to-human transmission was thought to take place either directly through physical contact with a sick person or indirectly by the clothes or bed sheets. In response, many towns and villages instituted quarantine regulations. The authorities in England, for example, recommended that plague victims be locked up in their homes or transferred to special "pest houses." An extreme example of adherence to public policy is the famous case of William Mompesson, rector in the small village of Eyam in Derbyshire, who persuaded the entire community to enter into quarantine when the plague erupted there in 1666. One by one the parishioners who remained faithful to their contaminated hearths fell victim to the disease. A mortality rate of 72 percent indicates that the community probably had a morbidity (infection) rate of 100 percent, an extraordinary price to pay for a misconceived theory.

Locking people in their homes is, of course, one of the worst possible ways to fight the plague. The plague is a disease of "locality," most likely to manifest itself when rats, fleas and people are kept in close contact with one another. To confine people is to maximize their chance of being bitten by a plague-carrying flea or infected through close contact with another human being.

Officials recognized that quarantines were dangerous to healthy individuals confined with sick relatives, but they imposed them nonetheless in the belief that some lives must be sacrificed in order to stop further spread of the disease. Because it is rats that carry the plague (and the rats were free to travel), the entire quarantine effort was a waste of time—and lives.

Attempts were also made to quarantine passengers and goods arriving in boats from overseas. When sickness suggestive of the plague was observed among crew members or passengers, the ships were diverted to lazarettos (quarantine stations) until the authorities deemed it safe to release them. At Marseilles in May of 1720, for example, the sailing ship *Grand Saint Antoine* was placed in quarantine for three weeks because eight of its crew had died in the course of the voyage from the Near



PROGRESSION OF THE PLAGUE through a medieval household could be very rapid once the black rats that lived there became infected. An infected rat, marked with a red dot on day 1, is shown to die from the disease by day 5. When that rat dies, its fleas leave, carrying the plague with them to other rats in the house. By day 10 these rats have also died from the disease and the fleas turn at that point to human beings, infecting almost 75 percent of them. By day 15 approximately half of the humans in the house will have died from the plague; a fourth of them will have recovered from it and a fourth will have escaped it.

East. In spite of these efforts to limit the spread of the plague, the disease broke out in Marseilles—first among the dockworkers who unloaded the ship's cargo when it was released from quarantine and then in the population at large.

There is little evidence that quarantines of this type were ever very effective. Venice was one of the first seaports to introduce quarantine regulations, early in the 15th century, enforcing them by imposing the death penalty on anyone who broke the rules. Yet Venice suffered from the plague as much as any city in Italy, presumably because it was impossible to prevent rats aboard quarantined ships from jumping ashore, carrying the plague with them.

Finally, after innumerable cycles of onslaught and retreat, the plague disappeared from Europe. London's last experience with the disease, the Great Plague, began in 1665 and ended in spectacular fashion with the Great Fire of 1666. At that time it was natural for Londoners to believe they owed their deliverance to the purifying conflagration. Later it was suggested Londoners owed their resistance to the plague to the reconstruction that followed the fire and the fact that the rebuilt city boasted brick houses and wide, rubbish-free streets in place of the higgledy-piggledy structures and malodorous alleys of medieval times.

This explanation is attractive but does not hold up under scrutiny. One reason is that the fire destroyed only the central part of London, the area least affected by any of the outbreaks of plague earlier in the century, leaving untouched the overcrowded suburbs that had provided the disease with its main lodging in previous times. A second reason is that other cities in Europe, such as Paris and Amsterdam, became plague-free during the same period—a phenomenon that could not be linked to the Great Fire of London.

A somewhat more convincing (but still flawed) theory suggests that the disappearance of the plague coincided with a slow rise in prevailing standards of health and hygiene. Although hygiene cannot be eliminated as a factor, it does not explain why subsequent outbreaks followed the standard course, complete with high rates of mortality, but were farther and farther away from the center of Europe each time they appeared. It was almost as if Europe were developing some form of resistance to the plague that kept the infection from propagating in the usual way. In the north the path of retreat was to the east; in the Mediterranean it was to the south. The later the epidemic, the less it seemed to be capable of spreading. This, moreover, was at a time when, according to every available index, traffic by land and by sea was increasing.

When the role of rats was finally established late in the 19th century, it was suggested that the subsidence of the plague could be explained by changes in the population dynamics of the black rat, *Rattus rattus*. During the 18th century it had been observed that the black rat, the historic carrier, had been largely displaced by a new species, the brown rat (*Rattus norvegicus*), which would have been a much poorer vector of the plague: the brown rat is as susceptible to the plague bacillus as the black rat but does not normally live in close proximity to humans. Brown rats typically live in dark cellars or sewers, whereas black rats overrun the upper rooms and rafters of a house. Because the oriental rat flea has a maximum jump of 90 millimeters (a little more than 3.5 inches), the difference in preferred habitats may have been enough to isolate humans from plague-infested fleas.

The brown-rat theory seems plausible but does not fit the geography: the brown rat spread across Europe in the 18th century from east to west, whereas the plague retreated from west to east. The brown rat was in Moscow long before the city experienced a particularly severe epidemic of the plague in the 1770's; it did not reach England until 1727, more than 60 years after that country's last bout of the plague.

The late Andrew B. Appleby of San Diego State University suggested an alternative theory, namely that a certain percentage of black rats became resistant to the plague over the course of the 17th century and that the resistant animals would have increased in number, spreading across Europe during the next 100 years. Although these rats might still be infected by the plague bacillus, they would not die from it and therefore could support a large population of fleas, rendering it unnecessary for the fleas to seek other hosts. This theory, however, does not conform to what is known about resistance to plague in animal populations. As Paul Slack of the University of Oxford has pointed out, rat populations often develop resistance when exposed to a pathogenic bacterium or virus, but such resistance is short-lived and is therefore unlikely to have been responsible for broad-based immunity to the plague.

A more plausible theory suggests that a new species of plague bacillus, *Yersinia pestis*, may have evolved that was less virulent than the previous strain. Being less virulent, it might have acted as a vaccine, conferring on infected animals and humans a relative immunity to more virulent strains of the bacterium.

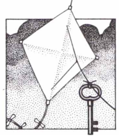
The bacteriological theory is acceptable on several grounds. First, it conforms to the dictum, proposed by the American pathologist Theobold Smith, that "pathological manifestations are only incidents in a developing parasitism," so that in the long run milder forms of disease tend to displace more virulent ones. Second, it explains why the decline of the plague is associated with a failure to spread beyond local outbreaks: a disease cannot travel far when the number of people susceptible to it is low. Third, it is supported by the existence of a close relative of the plague bacillus, *Yersinia pseudotuberculosis*, which does not induce visible illness in rats but does confer on them a high degree of immunity to the plague.

Did *Y. pseudotuberculosis*, or a relative with similar properties, gradually spread through the rodent population of early modern Europe, making it impossible for *Y. pestis* to gain a foothold there? Although no direct evidence exists to support that hypothesis, it seems more reasonable than any other.

The discovery and widespread use of antibiotics has conferred on human beings a different form of protection from the plague. Although the disease still occurs with regularity throughout parts of Africa, South America and the southwestern U.S. (in 1986, 10 cases were reported in the U.S.), it is never again likely to reach epidemic levels now that we know how it spreads, what public-health measures are appropriate and how to treat plague cases as they occur. Nevertheless, many questions about the plague are as yet unanswered. For example, the mode of transmission in rural areas, where rat populations are discontinuous, is entirely unclear. And what explains the distribution of the plague throughout the world today? Why are only certain rodent populations reservoirs of the disease whereas others are entirely free of it?

THE AMATEUR SCIENTIST

The feathery wake of a moving boat is a complex interference pattern



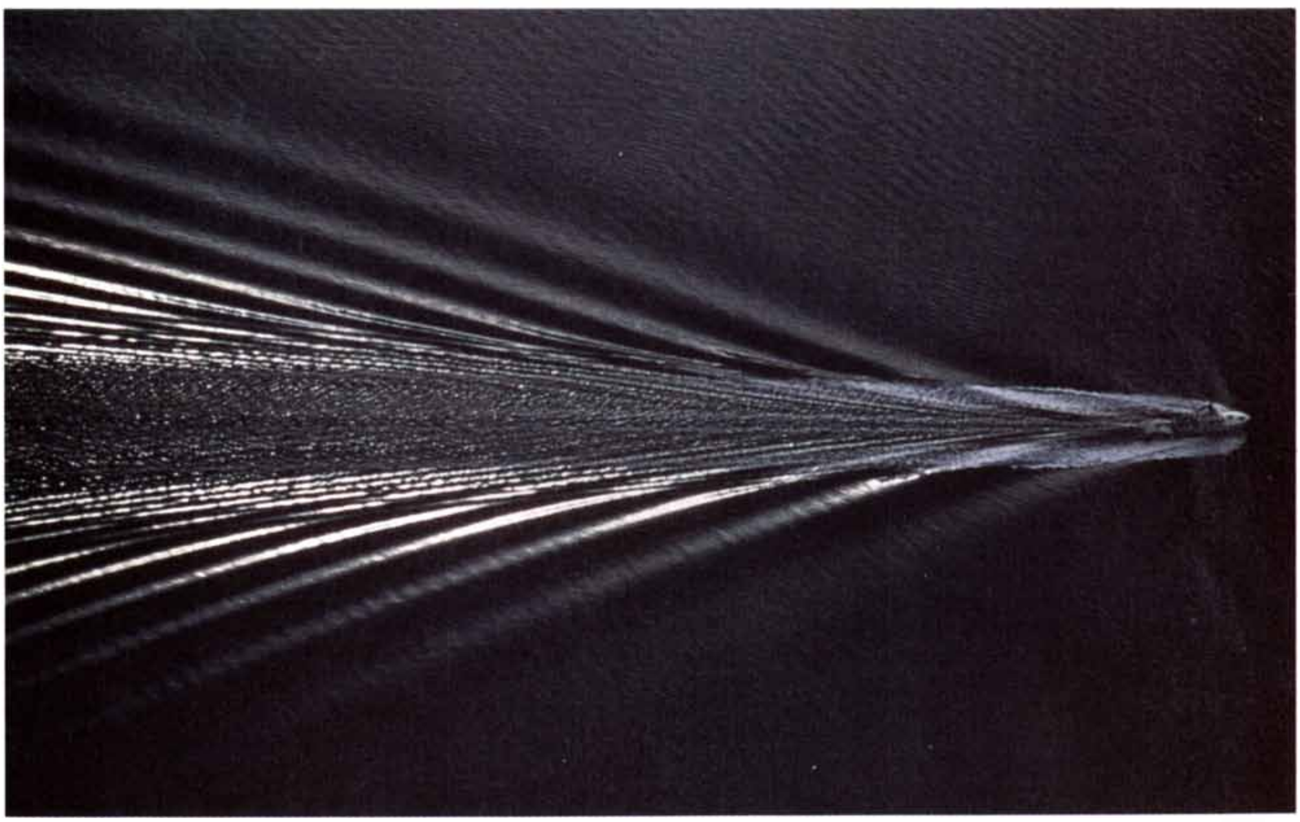
by Jearl Walker

Anyone who has flown over a waterway at a low or moderate altitude will have noticed the wake patterns left by moving boats. Each pattern is intricate, with a variety of waves contained within a V shape that travels along with the boat as if it were being towed. The general appearance of the wake may suggest that it is a shock wave similar to those shed by supersonic aircraft. (Although a shock wave in air is normally invisible, it can be photographed with special equipment when an aircraft model is mounted in a wind tunnel.) The shock wave also forms a V whose apex seems to be attached to the craft generating it.

In spite of the general resemblance of the two types of pattern, they are produced by different means. One clue to this fact lies in the angle between the arms of the V's. In a shock wave the angle depends on the relative speeds of the aircraft and of sound at the aircraft's altitude. If the aircraft's speed increases, the angle decreases, narrowing the V. Decreasing the speed widens the V, until it disappears when the plane's speed drops below the speed of sound.

If a wake pattern were only a shock wave, the angle of its V would also depend on speed. Actually the angle is always about 39 degrees, regardless of the boat's speed. As a matter of fact, anything at all, from ducks to giant petroleum tankers, creates a wake pattern with that same angle when it moves through water. Wakes also differ from shock waves in that they have a complex structure. The arms of the V are not single, long waves; instead they are made up of a series of short lengths of waves, which accounts for their feathery appearance [*see top illustration on opposite page*]. Extending between the arms there are curved transverse waves that seem to travel along with the boat.

The mechanism producing a boat's wake was first worked out by Lord Kelvin in 1887. The solution was quite challenging, requiring the invention of a mathematical tool called the method of stationary phase. The method enabled Kelvin to approximate how the waves created by a boat interfere with one another. Research into wake production



An aerial view of the wake pattern of a motorboat has continued to the present day, spurred by the fact that much of a ship's energy consumption goes to generating the wake pattern.

If you visualize a water wave, chances are that you imagine a sinusoidal wave traveling over the water surface at a certain speed. The wave, which is often called a phase wave, is characterized by crests and valleys that are evenly spaced and of uniform height and depth. The horizontal distance between adjacent crests is the wavelength. The frequency at which successive crests pass a given checkpoint is the wave's frequency, which is inversely proportional to the wavelength. The wave's amplitude is the height of a crest above a calm surface.

Although a phase wave is a convenient model, it is actually never seen on water because of a subtle property of water waves: their speed depends on their wavelength or (equivalently) their frequency. This fact guarantees that whenever you disturb a water surface, you see a complicated pattern, rather than a sinusoidal wave, spreading over the water. The pattern is the result of the interference of many phase waves, each traveling at a different speed; although individual phase waves are present, they are hidden from view.

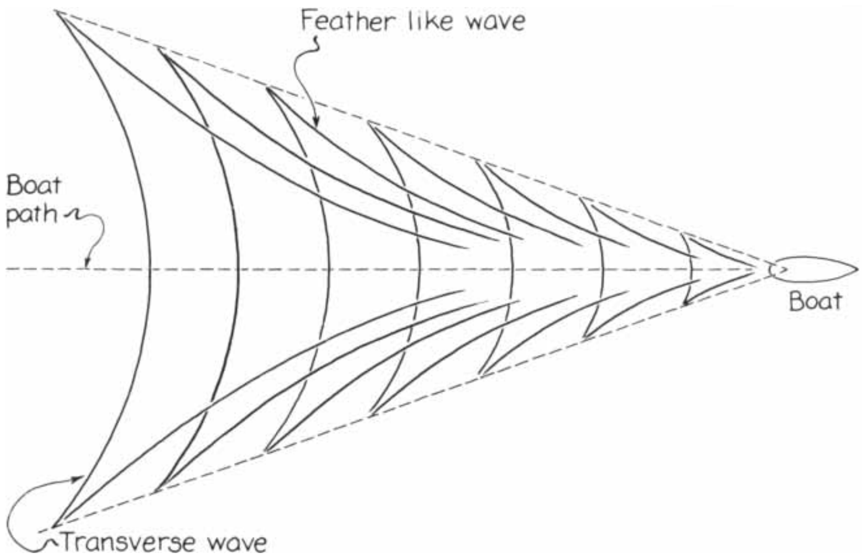
You can produce a water wave that is approximately a phase wave by arranging for a machine to produce periodic oscillations in the water at a certain frequency. (You must also eliminate the possibility that reflections will send waves back into the region you are examining.) If you then look at the resulting wave at some point that is not too near the machine, it is approximately sinusoidal. Its frequency matches that of the machine.

Suppose you measure the speed of the wave and then adjust the machine to generate a wave that has a higher frequency and therefore a shorter wavelength. You will find that the new wave travels slower than the preceding one. With enough experimentation you will discover that the speed of a phase wave depends on the square root of its wavelength. When the speed of a wave depends on the wavelength in any way, the wave is said to be dispersive. It is the dispersive property of water waves that shapes the wake left by a moving boat.

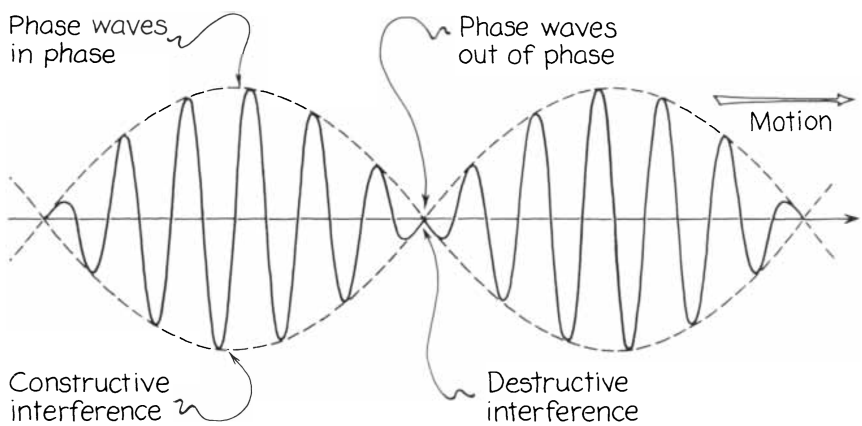
Textbooks often introduce dispersion with an example in which two phase waves differ only slightly in wavelength and wave speed. If the waves were sent over a water surface, they would interfere, giving rise to a moving pattern within which the phase waves are hidden. The wavy structure of the water surface is gathered into groups that move in a parade in the direction of the phase waves [*see illustration below*]. Crests and valleys progress through each group with twice the group's speed, appearing first at the rear of the group, growing larger toward the center of the group and then dying out when they reach the front.

The crests and valleys mark where the phase waves happen to be approximately in phase, or in step, so that they add constructively to form high crests and deep valleys. Toward each side of the group they are partially out of phase and interfere destructively, creating crests and valleys that are less pronounced. Between the groups the destructive interference is complete. Notice that although the groups seem to move as a whole (particularly if you are too far away to make out their internal structure), they are actually being recreated continuously as fresh sections of the phase waves progress through them. What you see move and what seem to persist are the regions of constructive interference.

If the two phase waves are replaced by phase waves that have a narrow range of wavelengths, the interference pattern is similar, with crests and valleys passing through the groups at twice the group speed. The pattern soon differs from the preceding example, however, because the phase waves with longer wavelengths outrun those with shorter wavelengths owing to dispersion. As the groups move away from the source of the waves, they change



The waves that make up a boat's wake



Groups formed by two phase waves

shape continuously. One way to describe the groups is as regions where the component phase waves are constantly in phase—hence Kelvin's term "constant phase."

Most disturbances of a water surface are more complicated than the preceding examples because they create phase waves that have a wide range of wavelengths. Suppose a stone falls into a large pond while you examine the water surface reasonably far from the fall. For a short time, perhaps tens of seconds, the disturbance sends out phase waves continuously. The ones with long wavelengths (and consequently with high speeds) are the first to begin passing you. With time, phase waves that move more and more slowly begin to pass you. Although there is soon a host of passing waves, the water surface remains calm because the waves all interfere destructively: every crest of a given phase wave is canceled by the valleys of other phase waves.

The earliest evidence of activity reaches you when the groups begin to pass. Each group travels at half the speed of its associated phase waves. The first group, the fastest, is produced by the phase waves having the longest wavelengths and therefore the highest speeds. With time, groups that are traveling ever more slowly go by, associated with phase waves that have progressively shorter wavelengths and lower speeds. Whereas in the first example there was a parade of identical groups, the stone gives rise to a succession of different groups that blend into one another.

The transition between groups is so smooth that you may have trouble seeing it, particularly since the action passes you quickly. If you photograph the display to freeze the action and measure the crest-to-crest separations, you will see that the faster groups are ahead of the slower ones. The separation is greatest in the region farthest from the impact of the stone, indicating that the phase waves with the longer wavelengths are responsible for that region. The separation gets progressively shorter as you examine the water surface nearer the impact point, indicating that phase waves with shorter wavelengths come into play there. By making a series of photographs you can see the groups stretch out as the faster phase waves race away from the slower ones, until energy losses and the spread of the waves finally eliminate all action.

A stone produces a single point disturbance. A moving boat creates a continuous series of point disturbances, each disturbance sending out expanding groups that interfere with one another to form the characteristic wake. Because Kelvin's solution to the interference problem is complex, I shall first give a much simpler account that was presented in 1984 by Frank S. Crawford, Jr., of the University of California at Berkeley.

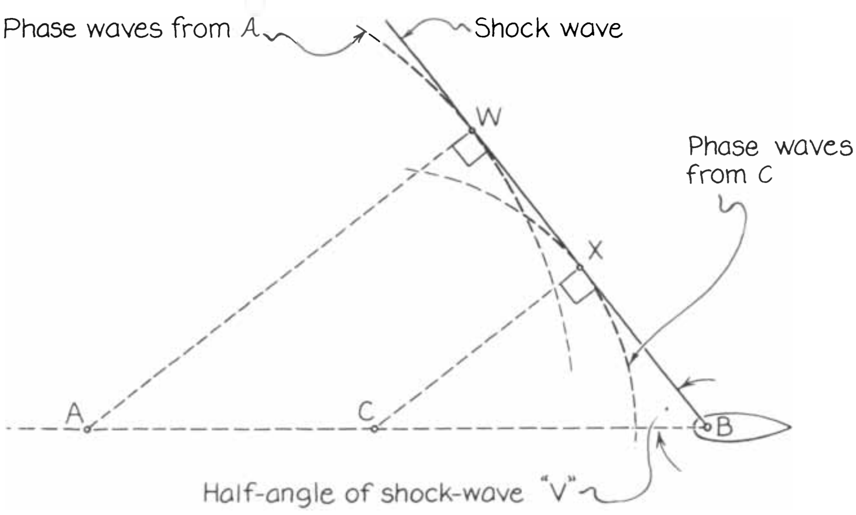
Crawford starts out by considering the shock wave a traveling boat would produce if water waves were not dispersive. In the illustration below a boat has traveled from *A* to *B* at a constant speed. Its disturbance at *A* sent out circular phase waves (all moving at the same speed, since we are ignoring dispersion). Phase waves were also created by the boat at each point between *A* and *B*, producing a composite pattern of circles of varying size, the boundaries of which form a V-shaped shock wave. By convention the angular extent of a shock wave is measured by the half-angle of the V. In the illustration the sine of that angle is equal to the ratio of the speed of the waves to the speed of the boat.

The fictional shock wave serves as a benchmark for constructing the actual wake created by dispersion. Although the boat's disturbance creates phase waves with a wide range of wavelengths, apparently the only waves that are important are those with speeds no greater than the boat's speed. First consider a small packet of phase waves that differ only slightly in wavelength and that travel at or very close to, say, .866 of the boat's speed. When the boat reaches *B*, the phase waves that were generated at *A* reach *W* on the arm of a fictional shock wave that, according to the equation, is at an angle of 60 degrees to the boat's path. Identical packets that were produced by disturbances between *A* and *B* (say at point *C*) also reach the arm of the fictional shock wave just then.

The shock wave is said to be fictional because the water along the arm is completely calm owing to the destructive interference of the phase waves there. The activity the phase waves do create is back in the groups that have traveled outward at half the speed of the phase waves [*see top illustration on opposite page*], and so the group associated with the phase waves reaching *W* is halfway between *A* and *W*. Draw a line between *B* and that halfway point. The line, which can be called the wake line, marks the current location of groups associated with the chosen phase waves that were generated at points between *A* and *B*.

If the boat had generated only these chosen phase waves, the wake would be contained by the wake line, and the wake angle, measured between the line and the boat's path, would be 19.1 degrees. The calculation of the wake angle goes as follows. Let *L* be the *AB* distance. Since the phase waves were chosen to travel at .866 times the boat speed, *W* must be .866 *L* from *A*, and the midpoint between *A* and *W* must be .433 *L* from *A*. Since the right triangle formed by *A*, *W* and *B* has one angle of 60 degrees, the angle *WAB* must be 30 degrees. The wake angle can then be computed by applying trigonometric rules to the triangle created by *A*, *B* and the midpoint of *AW*.

If you calculate the wake angles as-



The creation of a fictional shock wave

sociated with other speeds chosen for the phase waves, you will find that the wake angle has a maximum value of 19.5 degrees, which corresponds to phase waves that travel at .82 times the boat's speed. The wake line at this maximum angle marks the boundary of the pattern. It is most apparent in the pattern because a wide range of phase waves contribute groups on or just inside the line, generating prominent crests and valleys there. The full angle of the V, twice the maximum wake angle, is 39 degrees, which matches what is observed in actual wake patterns.

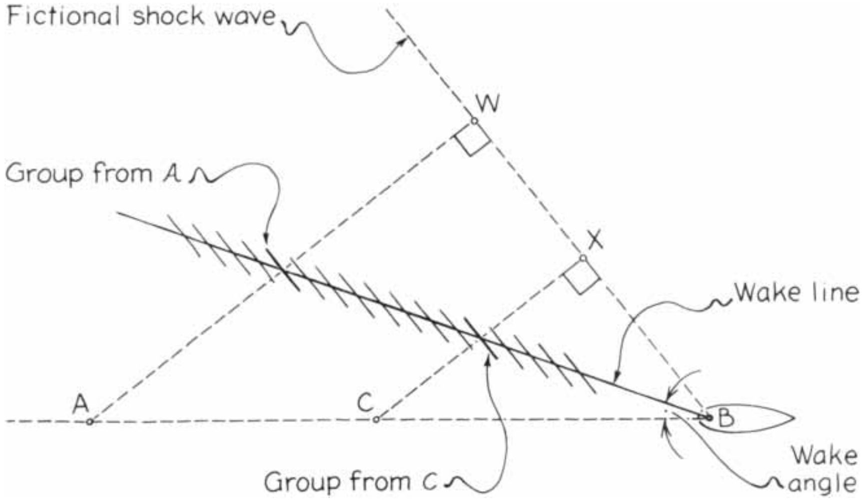
Crawford went on to consider the transverse waves in the wake pattern. They are due to the spreading out of groups toward the boat in circular arcs. Since the pattern remains stationary with respect to the boat, it must be dominated by a packet of phase waves that travel at, or only slightly above or below, the speed of the boat. The associated group, of course, travels at half the boat's speed. The packet created at *A*, then, creates a group that is halfway along *AB* when the boat reaches *B*. The other groups forming the transverse waves have a similar relation to the boat. If a group currently lies at a distance *x* behind the boat, it must have been created by a disturbance point that is now 2*x* behind the boat. Furthermore, the transverse wave that group forms must now have a radius of *x*.PThe composite pattern seems to be dragged along by the boat but is actually being re-created continuously by fresh phase waves as the transverse waves expand.

In 1957 James Johnston Stoker of New York University and Harvey Douglas Keith of the University of Bristol independently published simplified versions of Kelvin's analysis. In their versions the groups sent out from *A* are interfered with by groups sent out from other points. When the boat reaches *B*, the only surviving groups from *A* lie in a circle, with *A* on the circumference of the circle at the rear (with respect to the motion of the boat). The group at the front of the circle, produced by phase waves traveling at the speed of the boat, lies halfway between *A* and *B*. Other parts of the circle are dominated by groups associated with slower phase waves.

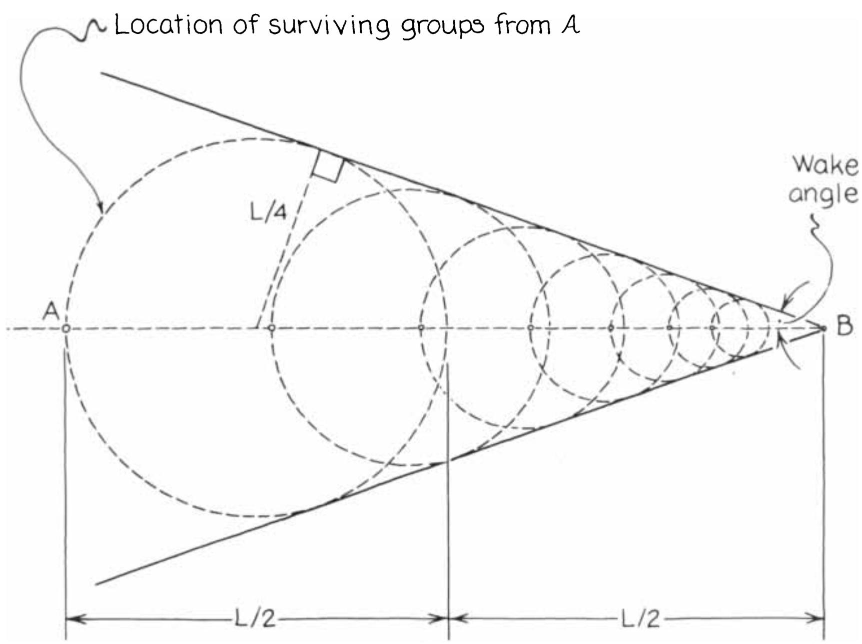
The circles are not waves; they merely indicate the location of the groups. To determine the orientation of a group at any point on a circle, draw a short line through the point in such a way that it is perpendicular to a line running from the point to the disturbance point associated with the circle. The short line represents the visible waves in the group at that point on the circle.

The illustration below shows several circles, each of which is linked to a disturbance point at the rear of the circle. Straight lines drawn along the sides of the circles define the outer limits of the wake pattern. One circle is chosen for a calculation of the angle of the lines—the wake angle. The disturbance point *A* at a distance *L* from the boat is on a circle that has a radius of *L*/4, the center of which is 3*L*/4 from the boat. Note that a radial line and the line from the center to the boat form part of a right triangle. The sine of the wake angle, which lies within the triangle, must then be equal to the ratio of the lines (1/3), and the angle itself must be 19 degrees 28 minutes.

I have considered the wake patterns for boats traveling in a straight line in deep water. How do the patterns change when the water is shallow or when its depth varies? What happens if the boat is traveling in a curved path? Why does the relative visibility of the transverse waves and the featherlike waves vary from one wake pattern to the next? I leave these matters for you to investigate the next time you fly over a waterway.



Groups for a particular packet of phase waves



Construction by means of Lord Kelvin's analysis

COMPUTER RECREATIONS

A blind watchmaker surveys the land of biomorphs

by A. K. Dewdney

If the mechanisms of life are likened to watches, then the watchmaker is blind, according to Richard Dawkins, Oxford biologist and author of *The Selfish Gene;* blind, physical forces drive evolution, after all. Dawkins has joined the debate between creationists and evolutionists on the latter side in a recent book called *The Blind Watchmaker.* To illustrate one of the book's central points, Dawkins has written a program that enables human beings to imitate the evolutionary process by designing their own graphic life forms, abstract creatures Dawkins calls biomorphs.

The program, which I shall simply dub WATCHMAKER, illuminates a fallacy often introduced by critics of the theory of evolution. How can something as complex as the machinery of life come about by the combination of chance events? On the face of it such a thing seems impossible. We would seek a watchmaker as surely as did William Paley, an 18th-century English theologian. Arguing for the existence of a Creator of all living forms, Paley imagined crossing a heath and stumbling on a stone. One might think it had lain there forever. "But suppose," Paley wrote, "I had found a watch upon the ground, and it should be inquired how the watch happened to be in that place; I should hardly think of the answer which I had before given....the watch must have had a maker."

Paley died before Charles Darwin was born. Creationists agree with Paley's argument, however. By what combination of chance events could a watch be assembled on Paley's heath? The fallacy lies in supposing the chance events leading to the watch (or its maker) happened more or less simultaneously. The argument does not take the effects of cumulative change into account. WATCHMAKER is a computational metaphor for the power of cumulative change. Its products are not living things but biomorphs: static forms that can resemble living things.

One begins with a very simple form in the center of a screen that is divided into large squares [*see top illustration on page 130*]. Suppose, for example, that the center square contains a small tree with a few branches. WATCHMAKER produces variant trees that fill surrounding squares. Some trees have more branches and some have less; some trees are tall and others are short. The way in which trees may vary is controlled by a number of genes. I shall return to these below. The variant trees are all regarded as progeny of the central tree. They represent the variations available in the current population.

WATCHMAKER runs on the Macintosh computer. Evolution is thus under the control of a "mouse"—the Macintosh mouse, that is. The mouse is an input device the user rolls across a pad by hand. The motion of the mouse on the pad is mimicked by a small, dark square on the Macintosh display screen. When WATCHMAKER runs, the user selects which tree to breed next by moving the mouse-controlled square into one of the adjacent variant tree boxes. A mouse has a button, and by clicking the button when the square rests in the selected box, the user causes that tree to become the ancestor of all subsequent trees. It migrates to the central box. The selection cycle begins anew. A sequence of selections generates a phylogenetic line of trees, each varying slightly from its parent. The slow and steady accumulation of small changes can have surprising results.

When he was about to run WATCHMAKER for the first time, Dawkins writes, "I had hoped for weeping willows, cedars of Lebanon, Lombardy poplars, seaweeds, perhaps deer antlers. Nothing in my biologist's intuition, nothing in my 20 years' experience of programming computers, and nothing in my wildest dreams, prepared me for what actually emerged on the screen. I can't remember exactly when in the sequence it first began to dawn on me that...something like an insect was possible.... My incredulity grew in parallel with the evolving resemblance." To the sounds of triumphal symphonic chords, distinctly insect-like creatures emerged [*see bottom illustration on page 130*]. Dawkins slept fitfully that night, his eyelids imprinted with swarms of insects.

So arose the first of the shapes that Dawkins eventually called by the generic term biomorph. Actually all shapes produced by WATCHMAKER are nothing more than trees. Branches that refold over the trunk or twist unexpectedly combine to produce not only the bodies, wings and legs of insects but also a myriad of other biomorphs, including tree frogs, bats and bee-flowers. Even technological shapes such as lamps and precision balances are possible [*see illustration on page 131*].

How does WATCHMAKER produce variations on a theme? As I stated earlier, the shape of each tree produced by the program is controlled by genes. There are 16 genes in all. The effects of some genes are easy to describe. The effects of other genes cannot be predicted because of interactions with still others. Thus some genes control the number of branches and overall size. Of the second type, three genes combine to control the horizontal extent of branches while another five jointly control vertical extent.

Dawkins conceives of a 16-dimensional space inhabited by every conceivable variation of biomorph. Each creature, after all, is determined completely by a raft of 16 numbers, its genetic endowment. The numbers can be regarded as coordinates, so that each biomorph becomes merely a point in what Dawkins calls biomorph land. WATCHMAKER is the vehicle of choice to explore this land.

Dawkin's program is too long and complicated to describe here. For one thing, it has a large number of options, each of which can be considered a program in its own right. For example, WATCHMAKER has an engineering option; users may read an accompanying manual to learn how to manipulate genes directly in order to produce variant biomorphs. Other options include a "fossil record" by which lines of descent can be stored like layered strata and a "hopeful monster" option that randomly selects successive progeny. It may be that something interesting, perhaps even a monster, will evolve.

By now the creationist will have cried "Foul!"—and fairly too. What sport is there in the unnatural selection of the surviving variant by a human? In reality this is nothing more nor less than deliberate breeding. Darwin used the success of human breeders to illustrate the existence of variation in populations. Of course, that did not prove natural selection; Darwin's wider argument depended on many sources, including fossil evidence. Dawkins' point has, of course, been made; the effects of cumulative change can be impressive. How much more impressive they would be if the cry of foul could be answered more directly. Rather than selection by humans or even random processes, could not WATCHMAKER's selection procedure somehow be put directly under the control of environmental pressures? Could morphology result from competitive interaction among species?

Dawkins has issued the following challenge to programmers: "Biomorphs should interact, in the computer, with a simulation of a hostile environment. Something about their shape should determine whether or not they survive in that environment." He goes on to say that vulnerability of organisms should emerge from the simulation itself and not be built in by the programmer. Even the number of genes might vary. Dawkins thinks if anyone is up to the task it would be the "programmers who develop those noisy and vulgar arcade games." Somehow I doubt it.

I cannot personally meet Dawkins' challenge, but I can illustrate the kind of situation he has in mind. It is, after all, possible to design ab initio two interacting species that change form as a result of their interaction. But vulnerability has been built in by me, the recreational WATCHMAKER. The following example might inspire some reader to leapfrog my modest progress into a truly evolutionary (not to say revolutionary) program.

Consider the bendosaurus and the spikophyte shown in the illustration at the right. The bendosaurus cannot survive without eating the succulent leaves on top of the spikophyte. Unfortunately the spikophyte is surrounded at its base by a collar of deadly spikes. As its name indicates, however, the bendosaurus has a bend in its body. The bend enables the bendosaurus to avoid the deadly points of the spikophyte while reaching its food.

That is the good news. The bad news is that the bendosaurus is dreadfully rigid. The angle of the bend is a permanent 90 degrees. Moreover, both neck and body are rigid. The bendosaurus is therefore committed to feeding on leaves at a certain level, provided its neck is long enough to reach them.

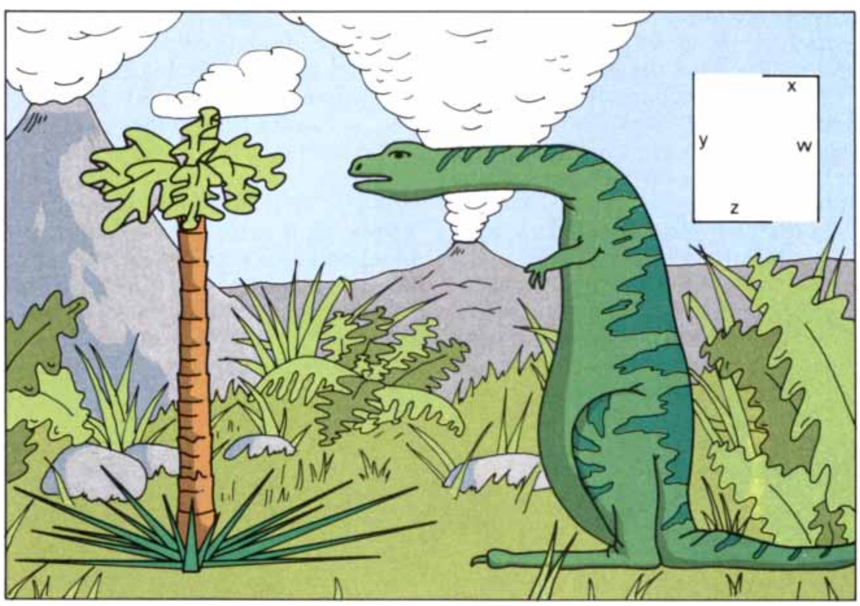
Naturally the populations of both creatures are subject to random variations in their genetic endowment. The bendosaurus can have varying heights (measured at the shoulders) and also varying neck lengths. The spikes of the spikophyte are all the same length in a given plant, but that length may vary from plant to plant. The height of the spikophyte is also subject to small, random changes.

As far as either biomorph is concerned, two genes suffice to control its form. The bendosaurus has a gene for tallness and one for neck length. The spikophyte has a gene for height and one for spike length. Under certain conditions embodied in the program I call ESCAPEMENT, the two life forms coevolve. This provides one part of a possible watch, if not its maker. In a long-term, genetic sense, the spikophyte seeks to escape the bendosaurus by evolving in a direction that makes it hard to eat.

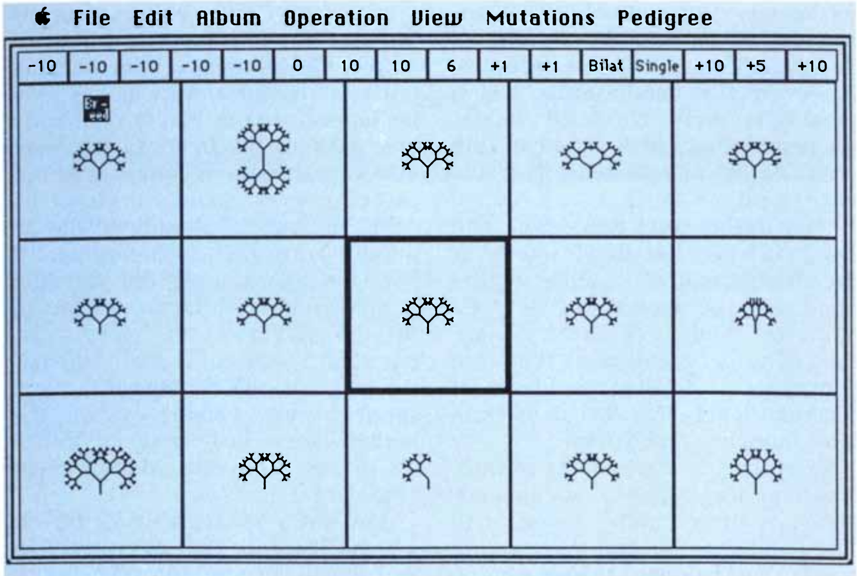
ESCAPEMENT gives both biomorphs turns at varying their form. For example, the spikophyte may become taller or shorter and its spikes may become shorter or longer. The body and neck of the bendosaurus may also vary, becoming longer or shorter. Because both biomorphs have absolute upper bounds on their size, an infinite arms race is impossible. It happens, for example, that the sum of spike length and trunk height in the spikophyte cannot exceed a certain limit I shall call *smax*. Similarly, the length of the bendosaurus, from snout to toe, cannot exceed the bound *bmax*. Just to make things interesting, I have decided to make *bmax* larger than *smax*.

How does ESCAPEMENT settle on which descendants of either biomorph will survive? The answer lies in a quantity called the feeding distance, which measures how far the bendosaurus's snout is from the fragrant leaves of the spikophyte. Suppose the bendosaurus has height w and a neck of length x. If the spikophyte has height y and spikes that are z units long, then the feeding distance is simply given by the Euclidean distance formula: the square root of the quantity $(z-x)^2+(y-w)^2$. The formula must be amended if z happens to be less than x. In this case the bendosaurus can stand as close as it wants to the spikophyte and only the second term is considered.

Not surprisingly, the spikophyte tends to evolve in a way that maximizes the value of the feeding distance. The bendosaurus, converse-



The spikophyte and the bendosaurus coevolve



Genetic variants of the central tree

ly, seeks to minimize the distance. When its turn comes, each biomorph gives rise to two variations. For example, the height of the spikophyte, currently *y*, may become $y-1$ or $y+1$ in the new generation. Spike length may increase from *z* to $z+1$ or decrease to $z-1$. In each case one of the two variations is chosen at random. Two variants of the bendosaurus arise in exactly the same way. Along with the variants, the parent stock is retained for evaluation against the biological opponent. Feeding distance provides the criterion. For example, whichever of the three spikophytes has the greatest feeding (or fed on) distance from the current model of bendosaurus is selected by the quasi-natural process and retained for the next evolutionary round. Now comes the turn of the bendosaurus. Its two variants and the parent stock are tried against the new spikophyte. Whichever has the smallest feeding distance becomes the proud new progenitor of its race.

It would be wonderful to provide ESCAPEMENT with the kind of graphic display typified by our artist's reconstruction of the two ancient and fabulous biomorphs. But that is beyond my capabilities. I have chosen instead the simplest kind of diagram, such as the one accompanying the illustration on the preceding page. Here little more than the relevant variables are represented by straight lines. Readers unwilling to attempt even this simple display must be content with having ESCAPEMENT print out the values of the four variables at the end of each round.

ESCAPEMENT employs a single loop within which both biomorphs are varied, evaluated and displayed:

input *bmax*, *smax*
input *w*, *x*, *y*, *z*
input *gen*
for $i \leftarrow 1$ to *gen*
  breed spikophyte variants
  evaluate spikophytes
  breed bendosaurus variants
  evaluate bendosaurs
  display winners

The user, of course, must input initial values of the six key variables by hand and also type in *gen*, the number of generations or rounds to be run. The loop then iterates two subcycles of breeding, evaluation and display repeatedly. A biomorph is bred by selecting two random numbers and using them to determine whether either gene will be turned up or down, so to speak. For example, the algorithmic code that generates variant bendosaurs produces mutations *a* and *b* of the genes *w* and *x* by the following technique:

$r \leftarrow$ random
if $r < .5$ then $a \leftarrow w-1$
  else $a \leftarrow w+1$
  $s \leftarrow$ random
  if $s < .5$ then $b \leftarrow x-1$
    else $b \leftarrow x+1$

The program now has three bendosaurs in hand. Their genetic endowments can be written as three pairs of numbers: $[w,x]$, $[a,x]$ and $[w,b]$. Which one is best? ESCAPEMENT evaluates the three models by computing their respective feeding distances from the current spikophyte. Dubbing the distances $d1$, $d2$ and $d3$, one can use essentially the same formula for each. Here, for example, is how ESCAPEMENT computes the second distance, $d2$, between the $[a,x]$ bendosaurus and the $[y,z]$ spikophyte:

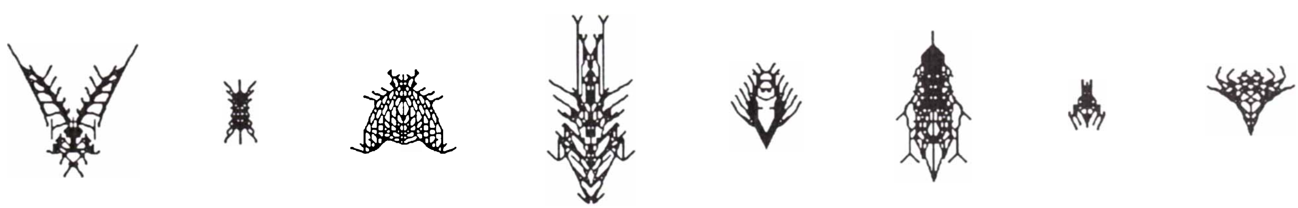
if $x < z$ then $d2 \leftarrow (z-x)^2 + (y-a)^2$
  else $d2 \leftarrow (y-a)^2$

In each case one must use the Euclidean distance formula with the appropriate mouth and leaf coordinates. When the three distances have been computed, the smallest is selected by a sequence of comparisons:

if $d2 \leq d1$ and $d2 \leq d3$
  then $w \leftarrow a$
if $d3 \leq d1$ and $d3 \leq d2$
  then $x \leftarrow b$

If neither of the two pairs of inequalities is satisfied, the parent continues to be superior to its progeny. The genes *w* and *x* remain unchanged. ESCAPEMENT evaluates spikophytes in an altogether similar way except that maximum rather than minimum distances are favored.

The foregoing description includes nothing about the growth limits I mentioned above. How does ESCAPEMENT prevent either biomorph from



Zoological biomorphs generated by WATCHMAKER

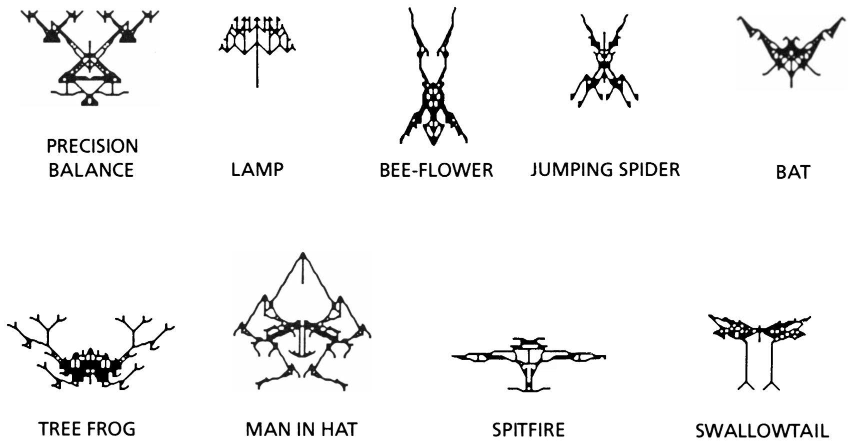
exceeding its inherent size limitation? The simplest way is to penalize the biomorph when the distance function is calculated. For example, if the variant bendosaurus [*a,x*] is too large, the sum $a+x$ will exceed the quantity *bmax*. In such a case 1,000 may be added to the distance just to prevent this variant from winning the evaluation process. Spikophytes are penalized by subtracting 1,000. Both creatures must be similarly penalized if any of the four variables becomes negative.

The winners of the evaluative procedures are displayed by the use of simple line-drawing commands of the kind commonly available in most programming languages. This part of the program is left to the imagination of readers.

What happens when ESCAPEMENT is run? That depends, in part, on the initial values of the four genes and two size limits fed to the program. Variables can also reflect screen size. Choose *smax* to be 100 and let *bmax* be 110. This choice gives rise to an interesting phenomenon. Initially, pick values for the four genes anywhere from 0 to 50. One may as well start with reasonably well-matched biomorphs, whether tall and narrow or short and wide, so to speak. As generation succeeds generation, a race ensues. The bendosaurus may "chase" the spikophyte into the tall, thin range of its variation, trapping it at the top. The spikophyte may then try to flee downward, thrusting out its nasty spikes in the process. The bendosaurus thereupon shortens and develops a longer neck. A seemingly endless cycle of flight and pursuit is guaranteed by the simple requirement that *bmax* exceed *smax*. Readers who have a good imagination might be able to visualize the result when the reverse is true.

Of course, the flight and pursuit mentioned above does not refer to individual biomorphs but to the fossil record. A paleontologist examining a complete sequence of strata containing fossil spikophytes and bendosaurs might be puzzled to see both forms covarying in size with time like the swinging of a pendulum.

The back of the paperback edition of *The Blind Watchmaker* contains an order form for the Macintosh program of the same title. There is a hefty discount, not surprisingly, if one buys the book. Readers unwilling to venture into the complex, subtle and engaging realm of Dawkins' biological exposition can order the program from the address above by noting on



Biomorphs and other forms generated by WATCHMAKER

the envelope "Attention of Department JDJ."

Dawkins would like to hear from readers who develop working experimental programs that explore the issue of emergent evolution. They could be elaborations of ESCAPEMENT or something much more profound. He can be reached by mail at the University of Oxford, Department of Zoology, South Parks Road, Oxford OX1 3PS, England.

Although the debate between creationists and evolutionists continues, one may wonder how much of the opposition on each side is falsely imagined. It is amusing to think of a sophisticated Creator, perfectly capable of creation through evolution, watching each side reducing its opponents to straw men.

The game called MAD (for mutually assured destruction) was featured in last October's column. Developed by Tad Homer-Dixon and Kevin Oliveau in 1985 at the Massachusetts Institute of Technology, the game simulates two superpowers poised between peace and war. In experiments at M.I.T. students took the command console at their computer stations, playing the role of chiefs of state. Should one cooperate by not going to war and accepting a modest increment in one's overall score or should one go to war, presumably to score a great many hit points? As the game progresses the potential rewards on the side of defection (striking first) pile up.

In spite of the fact that many students were told to maximize their total score, a goal that favors cooperation, many did not play this way. Nor did David I. Farmer of Mississauga, Ontario, who attempted to analyze a similar game by applying the principles of game theory. The theory treats only single game matrixes in isolation from subsequent moves. Game theory also seeks to maximize the total score by minimizing losses at each turn. In order to employ the theory, Farmer found a way to generate new game matrixes that reflect subsequent play. Farmer sent an example involving the five central matrixes of MAD. It makes sense to Farmer to go to war at the second matrix. But he concludes: "In the face of a growing threat of a devastating first strike it takes a rationale beyond game theory to keep the peace. If, however, the negative value of future war scenarios could be made less negative, even game theory would avoid the war option. The disarmament talks contribute to reducing the threat of a future war, thus modifying the payoffs in a positive direction."

In its last stages MAD is dominated by a situation known as the Prisoner's Dilemma. Here the payoffs for defection are so large that it seems to make sense to push the button. Yet a Prisoner's Dilemma tournament held several years ago at the University of Michigan revealed that tit for tat was the best strategy for the game; as long as your opponent cooperates, you should too. Now students (and others) may undertake their own tournaments featuring the Prisoner's Dilemma. Tournament software is available for the Macintosh and will soon be available for the IBM PC and PC-compatible computers from Leslie Burkholder and Chris Walton of Carnegie-Mellon University. Interested readers may inquire by writing to the authors at the Center for Design of Educational Computing, Carnegie-Mellon University, 5000 Forbes Avenue, Pittsburgh, Pa. 15213.

BOOKS

Dancing peacocks, a molecular field guide, futharks, a 1,000-jewel vegetable

by Philip Morrison

MONSOONS, edited by Jay S. Fein and Pamela L. Stephens. John Wiley & Sons, Inc. ($74.95).

The word is not just another name for rain; in Arabic *mausim* means an annual event, a season. For the first half of the year the days—although steadily getting warmer—are all the same, lucid and cloudless. The garden flowers bloom and wither; then it is the turn of the flowering trees to blossom into flaming scarlet and orange, until they too become seared. Finally relief from the now scorching heat is promised by the black wall advancing from the southern horizon: false hope, it is dry dust. The people are dejected, restless, sweating by night and by day. Then one morning there is a new dampness to the air; thunder is heard afar. Another black wall approaches swiftly, but this time it is cloud, glowing with the lightning of truth: rain, rain in big drops, rain in bursts, in waves, in teeming sheets. "All work stops. Men, women, and children run madly about the streets...shouting...hosannas to the miracle of the monsoon.... The summer monsoon is preceded by desolation; it brings with it the hopes of spring; it has the fullness of summer and the fulfillment of autumn all in one." Yet those all-nourishing rains may arrive late, be broken by dry spells or end prematurely. Pitiless drought and irresistible flood alike (20 inches of rain in one day is not rare) come and go to the irregular beat of the monsoon.

The monsoon climate is an annual wind reversal that marks two tropical seasons. The classical signal of the change is the soaking-wet summer wind just described. Each June it begins to blow stiffly out of the ocean to the southwest and then drenches most of India with 100 days of rain. Across the big peninsula a complementary rhythm, the winter monsoon, blows in from the northeast late in the year; briefer, it supplies half of the rainfall of southern India. Southeast Asia, Indonesia, Australia and West Africa similarly lie under well-marked monsoons; perhaps half of the land between and near the Tropics is watered by these alternations.

The Washington editors of this volume (at the National Science Foundation) have assembled a score of chapters by scholars and scientists around the world to set before the general reader a many-sided account of the monsoon. The focus is on Asian monsoons, in particular the Indian summer monsoon. An opening chapter outlines the introductory physics of the grand rhythm in a cleverly simplified model, expanded in a later chapter to confront the full theories. Several chapters treat of the monsoon in culture; they present Sumerian references, the Sanskrit writers and today's popular sayings as well as high art. The historical record is mustered to survey the actual monsoons of the past, particularly those reported by the Arab sailors and scholars who learned to navigate through them. There is a chapter on how modern science came to grasp the nature of the monsoon, and a series of deeper technical accounts that explain the meteorology and oceanography underlying variations in the intensity of the storms. The last set of essays are statistics-rich dispatches from the meteorological front lines by modelers and forecasters who try to assess the rains to come, both in the short term and the long. The volume ends with a concise proposal for managing the inconstancy of such weather as forecasts grow better.

The monsoon is more complicated than what its pioneer theorist, Edmund Halley, saw in the early 18th century: a form of the familiar fair-day onshore breeze that blows cool from the sea by day and reverses at night. Although that begins to model the realities, the monsoon exists at a wholly larger scale. Indeed, it involves another substance, water: 1,000-mile gyres forced by the earth's rotation. From their warm tropical surfaces vapor enters the rising, cooling air. Its latent heat accounts for the intensity of the monsoon. The storm is actually triggered by air rising over hot continental land. As the rains fall the lands below become wet, more oceanlike; the rain front migrates inland as the days pass, until the coastal strip has time to dry again. The breaks in the rains are built in. High aloft the winds flow back to the ocean.

A full model in three dimensions with all the topography is what we can begin to use today in the Age of Silicon. It is no surprise that computations suggest the Tibetan highlands augment the summer heating of the upper air over India. Even 10 years ago a benchtop analogue experiment, done in China with rotating rings of glycerin-thickened water, sources of heat and cold and a hand-size block to model the Himalayan rise, could show how summer flow was affected by that great plateau.

But most remarkable is an unlikely effect that offers new promise for reliable forecast. It was uncovered at the opening of the century by Sir Gilbert Walker, director of the Indian Meteorological Department. There were hints before then, but by 1923 Sir Gilbert had put together enough data to point to "a swaying of pressure on a big scale backwards and forwards" along the Equator, a barometric seesaw between the Indian Ocean and the shores of South America. By now the Southern Oscillation is well recognized. Every once in a while the coldest area of all the tropical oceans warms up, just off Peru. The ocean currents reverse, fish die, extraordinary Christmastime rains soak the arid desert nearby and even far-off cloudy Seattle becomes sunny. They call it El Niño, the Christ Child. During such years the Indian monsoon is likely to be scant.

It is still not quite clear how this works, and yet the modelers seem close to effective understanding. The initial steps were taken by Jacob Bjerknes more than 20 years ago. He saw how a dynamical ocean-wind connection could explain Walker's statistical global teleconnections. The long pressure waves interact with the sea surface; there can be strong feedback between the motions of sea and air. Abnormal sea currents from the west flow straight along the Equator (there alone the rotation of the earth has no deviating effect), pushing a tongue of warmer surface water from the hot Indonesian ocean eastward so that it overlies the cool Peruvian waters. Somehow a barrier of winds is also set up far westward above the Arabian Sea. The barrier impedes the cross-equatorial flow that yearly brings wringing-wet air from the warm southern Indian Ocean to feed the summer monsoon. That was the state of knowledge in 1984. Newer models, newer data from ship and satellite and new hope for timely forecast of the most important meteorological syncopation on the earth seem just ahead.

El Niño has been uncommonly strong a couple of times in the last decade. Indeed, the summer monsoon of 1987 was again seriously inadequate. The crop loss in India was covered by the 20-odd million tons of grain prudently stored over the past few years by the Republic of India. The matter is no light one; many poor families are in current want, although not yet in the grip of famine. Tube wells moderate the waywardness of the rains by watering most winter wheat fields. Wheat crops remain steady now, but a bad monsoon cuts notches as deep as a third of the crop into the rising curve of rice harvests. The supercomputers and the models that animate them have practical impact. Payments for public works to improve productivity in good years, like grants of food in bad years for the nutrition of the most imperiled, are policies that can be implemented more effectively if there is foreknowledge.

One day the impersonally variable monsoon will be rightly received with the unstinted praise of the sixth-century poet Subhandu: "Peacocks danced at eventide.... Lightning shone like a bejeweled boat of love in the pleasure-pool of the sky.... The rain was like a chess player, while yellow and green frogs were like chessmen jumping in the...fields.... By and by, the rainy season yielded to autumn, the season of bright dawns; of parrots rummaging among rice-stalks; of fugitive clouds."

MOLECULES, by P. W. Atkins. Scientific American Library, distributed by W. H. Freeman and Company ($32.95).

Dozens of rightly popular field guides catalogue and picture the diverse species of birds, wildflowers and minerals. Here is a field guide of a much more abstract kind, done with the brilliance, liveliness and richness of visual image familiar to the admirers of Roger Tory Peterson. The species it presents, however, are far more numerous even than sparrows or dandelions.

"In fact, you are made of molecules." The intention of this book is to unpack that truism, to strengthen and ornament it with rich example so as to make the molecular models visually familiar. It is conceded that molecules are not generally perceptible, however universal their presence in the stream of daily experience; we see molecules not one by one but in great unresolved flocks.

P. W. Atkins, an Oxford physical chemist and a writer of sensitivity, wit and taste, has made his field guide around standard maps—actually they are color-coded, drawn scale models of clustered atoms—representing some 160 distinct molecules. These are the primitives of this discourse; the scale of the map is as distinct from real size as a state highway map differs from the real lands, although the patterns are in some ways faithful. This terrain has been mapped and given explicit sculptural form both through the tested inferences and much extended senses of three centuries of modern science and through phenomena the reader can perceive directly. Indeed, it is possible to "comprehend the kingdom I will show with immediacy, without prerequisites, without labor, with a pleasure akin to looking at a work of art."

The author is as good as his word; the pages for every molecule whose form is shown provide reading as clear as that found on the page in a book that tells the bird watcher about the house finch, its distribution, habits, importance and relatives. All these rings, clumps, lumpy columns and broad repetitive webs are built according to the same sculptural precepts, mainly out of eight common kinds of atoms. The mineral kingdom is essentially absent: this is the traditionally chemical view of matter as consisting of finite molecules. Take that view with a grain of salt; the very saltshaker you might use is absent—no crystals, no glass, no metal cap. But what is here fascinates: "With luck, you will see a little through the mist and understand how a chemist thinks."

Octane is one molecule with good name recognition. The straight chain of eight carbons with their hydrogen encrustation is a simple linear array.

Venus and Mars, ***by Botticelli, in the National Gallery, London***

Gasoline is a mixture. Most of its molecules have about the same number of linked atoms as octane has but many bear branded hydrocarbon side chains. Octane is not really straight; rather, it is a zigzag linking, twisting easily at every joint. In any sample of the liquid the tiniest droplet contains more molecules than there are stars in the galaxy. Some molecules are rolled into a tight ball, some are stretched out, most are intermediate. Nor is the molecule ever stationary, as it has been drawn on the page. "Octane molecules are constantly writhing and twisting, rolling and unrolling, so that a gallon of gasoline is more like a can of molecular maggots than a box of short sticks."

Such insightful detail is presented for a wide variety of molecules. The guide opens with simple ones, such as the single unlinked atoms in the air of every breath: there are atoms of elemental argon, still a chemical virgin. Atkins goes on to describe molecules whose names are well-worn acronyms, such as ATP, the fuel of cells. You will look at the structural polymers made by humankind and by nature, the specific substances that activate the subtle chemical sensors of taste and smell, the molecules that allow vision and those that stain the living world with color, and finally a few of the ones that determine our lives at depth, the drugs and hormones.

Here is the ancient intoxicant ethanol, with a simple account of its role in binding to and distorting the "upstream" protein surface of a synapse, or nerve-cell junction, of the brain. Later we see the symmetrical molecule of TNT, its six oxygen atoms close to the brink of oxidation; a mere nudge and the atomic reaction begins, a swift inward release of energy, a submolecular explosion, not just the slow combustion that takes place when enough time elapses for oxygen atoms in the air to come in contact with the fuel molecules of coal or gas or oil.

The hydrogen bond is a repeated motif. A fine set of pages outline the sugars and their polymers, from the simplest, glucose, to the long polymer chains of starch and cellulose. The difference between starch the fuel and cellulose the scaffolding is the mere twist of a link, "unconscious nature at its most brilliant."

One photograph shows a scatter of blue cornflowers; below it another displays a bright border of scarlet poppies. The pigments that color the reflected light from the two kinds of blossom differ by a single atom in a cluster of about 30. The colorant is the violet indicator dye cyanidin, abundant in blackberries. The acidic sap of the cornflower is rich in hydrogen ions. The cyanidin molecule acquires one in just the right niche; the loaded molecule looks blue, since it absorbs reddish light efficiently. In the alkaline juices of the poppy few free hydrogen ions are present, and that molecular niche of the dyestuff stays vacant. In the poppy blossom the pigment reflects red light.

A more telling example yet of the profound difference a minor molecular change can make within living systems closes the book. The two molecules that are compared belong to the class known as steroids. The steroid testosterone, a key male sex hormone secreted by the testes, is a curtailed version of cholesterol: it consists of a chunky array of several dozen atoms built into a few fused rings. Replace with an $-OH$ pair one oxygen that lies off at the end, and delete a $-CH_3$ at one important carbon corner. That quite similar array is estradiol; it is one of the principal female sex hormones, abundant in women from puberty to menopause. The dramatic difference a methyl group can make is shared with us mortals by the gods themselves. Atkins reproduces the wonderful Botticelli painting that shows weary Mars asleep beside Venus pensive. Her cherubs, deficient in steroids, are at play with the casque and lance the war god has set aside.

The book opens with a Miró as apt as the old Florentine master with which it closes. Never too technical, full of matter to delight eye and mind, the work outlines an easy road along which any attentive reader can stroll, to share a little of the viewpoint and aesthetic of the reflective chemist.

RUNES, by R. I. Page. University of California Press/British Museum (paperbound, $6.95).

The old Germanic rune row is known as the futhark. Certainly it cannot properly be called an alphabet. Its 24 runes conventionally begin with the six signs of the word (among which a single rune represents *th,* the initial sound of *think*). The number of runes, the similarity of half a dozen of them to Roman (or Greek) letters, the dates and places of early inscriptions—all leave little doubt that the script originated among Germanic peoples along the frontiers of imperial Rome. A century or so elapsed from that time until the oldest inscriptions we know were rendered, no later than A.D. 200, in southern Denmark. Beyond that general assignment, with its implication that runes are an adaptation of Roman writing for the purposes of the barbarians, we cannot now go. Runes might have been first devised in Denmark or eastward on the Danube or southward in the Swiss Alps. Whatever their origins, they have been used in varied forms for 1,500 years to record at least a dozen Germanic tongues including Gothic, Scandinavian, Frisian and Anglo-Saxon.

Runes were to a degree wonders in that rude world without books. A gilt brooch might bring luck if it bore the rune list. By the ninth century there are manuscripts in alphabetic script that list the runes, give all of them names accompanied by meaning and try to tell us the sound value each expressed. The names, reconstructed with much uncertainty, are vivid enough: monster, sun, hail, gift, birch twig, horse and so on. One page in this pleasing little bargain of a book bears a photograph of 23 wood slips excavated at a site in Bergen in western Norway, datable to the 12th century. They are tallies, each bearing the name of a merchant, meant to be tied to or stuck into packages of goods.

Runes are linear marks designed for easy incision into wood. They are formed of strokes that would stand out against the grain; many have slanting side arms on one or two vertical strokes cut straight across the grain. Horizontal marks were avoided. In those times everyone carried a knife, and the work of shaving a flat area or two on any handy stick was quick and easy.

This is no medium for books, even though runic inscriptions soon spread to all kinds of surfaces. You can see them here incised or die-stamped on gold sheet, on coins and medallions, on a celebrated jewel box of whalebone, on clay pots and iron shields, on tablets and boulders. Many texts are transcribed here, and their import is discussed. We read of Viking runic graffiti carved into Saint Sophia in Istanbul and on a marble lion from the Piraeus. These sparse texts, sometimes consisting of a single word, are hard to construe reliably. An early spearhead bears one word that is taken as being a name, but it is read by some scholars from left to right and by others from right to left. The order was in fact not fixed for runic text; what credence, Professor Page asks, should readers give to linguists who can manage some sense out of either order?

The author writes like that, lightly, clearly and candidly, with learning and reason. Somehow the epigraphy of runes has with time harbored a long string of romantics. The runes are professed to be a cult script, meant chiefly for magic, as it is said rune names show. Certainly there are magical inscriptions, and amulets exist. A few stones even bear explicit curses and hints of magic. But when the rune master is the only literate person, how else could magical texts be written?

Rune stones are claimed from some 40 sites in North America. Many show some linear marks that may—or may not—be runes. A few offer brief but detailed texts in Norse by Vikings here in the New World. One slab found in Maine in 1971 even has cut into it a map of Vinland. "If you look for something single-mindedly enough, you are likely to find it; or at any rate something that looks like it; or at least something that has been made to look like it." Not one American find appears to be both genuine and runic.

The corpus of runic texts is some 5,000 inscriptions, most of them from Sweden, where medieval rune stones are common, erected to bear public witness, usually as memorials. The earlier texts provide insight into Viking life from the side of the tough Vikings themselves, to set next to the outraged chronicles of their victims overseas. Many inscriptions commemorate prowess in battle, some are even self-glorifying. But not all. A graceful pillar once marked a grave mound on a farm in Norway. It bears a design, the Wise Men under the Christmas star. The loving text carries the resonance of a society where men might often die far from home, with familial responsibility devolving on women: "Gunnvǫr, Thrýðrík's daughter, made a bridge in memory of her daughter Ástríðr. She was the most skillful girl in Hadeland."

This book is one of a new series called Reading the Past; their price, quality and style are exceptional.

The Best of Japan. Kodansha Ltd., distributed in the U.S. by Harper & Row, Publishers, Inc. (paperbound, $24.95).

Just 20 years ago there appeared a volume of eye-catching photographs titled after the stunning way the farmers of northern Japan wrap five eggs in a few lengths of artfully twined rice straw. A couple of hundred other examples of traditional Japanese packaging design were joined to that one, documenting how the universal sensibilities of the old, distinct island culture are expressed even in the most commonplace and ephemeral undertakings of life. Eggs are perhaps still carried home in such a wrapping, but today's Japan is seen as a tireless source of innovative wares of every kind. The wares command the admiration of millions of customers around the world, not that of a handful of neighbors of a farmer-craftsman who carefully prepared his product in the one way that had long seemed exactly right.

This book shows us some 250 commercial products of Japanese design, all of them made within the past five years, each accompanied by a brief account of its nature, development and prospects. They were chosen each year by an expert panel as leading examples of creative excellence among new products and services. Nine academics in engineering and commerce judge on behalf of the economic newspaper *Nikkei.* About half of the book is devoted to the 1986 examples, which are celebrated in color photographs and longish summaries. The awards span a remarkable diversity: they include a hand-operated toy cat's paw able to show a catlike way of beckoning as well as a liquid-oxygen and hydrogen rocket engine of 10 tons' thrust.

No five-egg wrap is explicit here, but the same sense of craft, the grasp of detail, the ingenuity of solution and the patient mastery of a process are implicit. This is a breathlessly up-to-date mixture of powerful high-tech, eccentric fashion, concern for the everyday and avid salesmanship. The aesthetic strength of the old ways is only rarely seen, but it seems likely that the past was not so different. The superb packages remain; the others are hidden from us by all those dusty years.

Look at a few products. The first is a miniaturized component of assured success. It is the joint accomplishment of two powerful electronics firms, Sony and Sharp. The laser diode, a small cylinder with leads two millimeters apart, is the infrared source and detector that illumines and scans an audio compact disc. No single part is more important to CD-player design: the smaller the diode, the smaller the player itself can be. The size makes portable players possible. The first IR semiconductor lasers at room temperature were a 1970 development of the AT&T Bell Laboratories. It took the Japanese firms years to improve the working lifetime of such diodes; the norm is now 20,000 playing hours. Japanese manufacturers today have a 99 percent share of the world market. The new version, a fifth the size of the older forms, delicately assembled by automated means, was put into production in mid-1986. Half a million diodes were being made each month by the beginning of 1987. The laser chip is critically aligned within the $15 assembly. "It appears that...Sharp and Sony laser diodes will be the world standard."

Ice cream and other frozen desserts have long been worldwide pleasures. Their production at home has never been easy. Even where the refrigerator is ubiquitous, ice cream of quality requires more; it must be stirred during the freezing process. An external freezer works well, but it demands an ice supply. A separate refrigerator is heavy and expensive. Simple motor stirrers that turn while the container is hidden within the freezing compartment are old but not very successful. In 1983 Nippon Light Metal found a solution as elegant as any egg wrap, if less beautiful: their container is a big double-walled cup within which a refrigerant is sealed. Seven hours in any freezer solidifies the refrigerant. At low temperature the solid substance has enough heat capacity to freeze any dessert mix in 15 minutes while the ice cream is stirred by hand at the table, in full view and control. It is a brilliant and frugal solution, exactly suited to the circumstances.

There is more, much more: a paper-thin piezoelectrical loudspeaker; three-dimensional fabrics, strong in all directions, woven of fibers such as graphite and silicon carbide, now entering mass production; a new hybrid—the 1,000-jewel vegetable—common cabbage crossed with a Japanese spinachlike plant (the new form was first grown by embryo tissue-culture techniques until seeds could be produced); meter-size shock mounts for buildings, many laminations of steel and rubber able to bear a couple of hundred tons, fit to absorb earthquake shock over their life expectancy of 60 years; seal-bearing credit cards...

It is forgivable hyperbole to call these commercial products, fine as they are, the best of Japan. But the book provides a view of the market that illuminates economic innovation in the world today, and Japan's high place within it.

BIBLIOGRAPHY

Readers interested in further exploration of the subjects covered by the articles in this issue may find the following lists of publications helpful.

CHROMOSOME MAPPING WITH DNA MARKERS

MULTILOCUS LINKAGE ANALYSIS IN HUMANS: DETECTION OF LINKAGE AND ESTIMATION OF RECOMBINATION. G. M. Lathrop, J.-M. Lalouel, C. Julier and J. Ott in *The American Journal of Human Genetics,* Vol. 37, No. 3, pages 482-498; May, 1985.

CONSTRUCTION OF HUMAN GENETIC LINKAGE MAPS, I: PROGRESS AND PERSPECTIVES. R. White, M. Leppert, P. O'Connell, Y. Nakamura, C. Julier, S. Woodward, A. Silva, R. Wolff, M. Lathrop and J.-M. Lalouel in *Cold Spring Harbor Symposia on Quantitative Biology,* Vol. 51, pages 29-38; 1986.

INVESTIGATION OF GENETIC LINKAGE IN HUMAN FAMILIES. Ray White and Jean-Marc Lalouel in *Advances in Human Genetics,* Vol. 16, pages 121-228; 1987.

VARIABLE NUMBER OF TANDEM REPEAT (VNTR) MARKERS FOR HUMAN GENE MAPPING. Yusuke Nakamura, Mark Leppert, Peter O'Connell, Roger Wolff, Tom Holm, Melanie Culver, Cindy Martin, Esther Fujimoto, Mark Hoff, Erika Kumlin and Ray White in *Science,* Vol. 235, No. 4796, pages 1616-1622; May 27, 1987.

A FLAW IN A UNIVERSAL MIRROR

EFFECTIVE HAMILTONIAN FOR $\Delta S=1$ WEAK NONLEPTONIC DECAYS IN THE SIX-QUARK MODEL. Frederick J. Gilman and Mark B. Wise in *Physical Review D,* Vol. 20, pages 2392-2410; November 1, 1979.

MEASUREMENT OF THE *CP*-NONCONSERVATION PARAMETER ϵ'/ϵ. J. K. Black, S. R. Blatt, M. K. Campbell, H. Kasha, M. Mannelli, M. P. Schmidt, C. B. Schwarz, R. K. Adair, R. C. Larsen, L. B. Leipuner and W. M. Morse in *Physical Review Letters,* Vol. 54, No. 15, pages 1628-1630; April 15, 1985.

THE GREAT DESIGN: PARTICLES, FIELDS, AND CREATION. Robert Adair. Oxford University Press, 1987.

GABAERGIC NEURONS

EFFECTS OF PICROTOXIN AND STRYCHNINE ON RABBIT RETINAL GANGLION CELLS: LATERAL INTERACTIONS FOR CELLS WITH MORE COMPLEX RECEPTIVE FIELDS. J. H. Caldwell, N. W. Daw and H. J. Wyatt in *The Journal of Physiology,* Vol. 276, pages 277-298; 1978.

MONOCLONAL ANTIBODIES TO GLUTAMIC ACID DECARBOXYLASE. David I. Gottlieb, Yen-Chung Chang and James E. Schwob in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 83, No. 22, pages 8808-8812; November, 1986.

SEQUENCE AND FUNCTIONAL EXPRESSION OF THE $GABA_A$ RECEPTOR SHOWS A LIGAND-GATED RECEPTOR SUPERFAMILY. Peter R. Schofield, Mark G. Darlison, Norihisa Fujita, David R. Burt, F. Anne Stephenson, Henry Rodriguez, Lucy M. Rhee, J. Ramachandran, Vincenzine Reale, Thora A. Glencorse, Peter H. Seeburg and Eric A. Barnard in *Nature,* Vol. 328, No. 6127, pages 221-227; July 16, 1987.

HOW CLIMATE EVOLVED ON THE TERRESTRIAL PLANETS

ATMOSPHERES. Richard M. Goody and James C. G. Walker. Prentice-Hall, Inc., 1972.

CLIMATIC CHANGE ON THE TERRESTRIAL PLANETS. James B. Pollack in *Icarus,* Vol. 37, No. 3, pages 479-553; March, 1979.

A NEGATIVE FEEDBACK MECHANISM FOR THE LONG-TERM STABILIZATION OF EARTH'S SURFACE TEMPERATURE. James C. G. Walker, P. B. Hays and J. F. Kasting in *Journal of Geophysical Research,* Vol. 86, No. C10, pages 9776-9782; October 20, 1981.

THEORETICAL CONSTRAINTS ON OXYGEN AND CARBON DIOXIDE CONCENTRATIONS IN THE PRECAMBRIAN ATMOSPHERE. James F. Kasting in *Precambrian Research,* Vol. 34, Nos. 3-4, pages 205-229; January, 1987.

THE ADAPTABLE OPOSSUM

THE BIOLOGY OF SENESCENCE. A. Comfort. Elsevier North-Holland Publishing Co., 1979.

VIRGINIA OPOSSUM. Alfred L. Gardner in *Wild Mammals of North America: Biology, Management, and Economics,* edited by Joseph A. Chapman and George A. Feldhamer. The Johns Hopkins University Press, 1982.

SEX RATIO VARIATION IN MAMMALS. T. H. Clutton-Brock and G. R. Iason in *Quarterly Review of Biology,* Vol. 61, No. 3, pages 339-374; September, 1986.

PLASTICS THAT CONDUCT ELECTRICITY

POLYACETYLENE: CHEMISTRY, PHYSICS, AND MATERIAL SCIENCE. James C. W. Chien. Academic Press, 1984.

HANDBOOK OF CONDUCTING POLYMERS. Terje A. Skotheim. Marcel Dekker, Inc., 1986.

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SCIENCE AND TECHNOLOGY OF SYNTHETIC METALS. In *Synthetic Metals,* Vol. 17-19; January-March, 1987.

RAMANUJAN AND PI

MODULAR EQUATIONS AND APPROXIMATIONS TO π. S. Ramanujan in *The Quarterly Journal of Pure and Applied Mathematics,* Vol. 45, pages 350-372; 1914.

COMPUTATION OF π USING ARITHMETIC-GEOMETRIC MEAN. Eugene Salamin in *Mathematics of Computation,* Vol. 30, No. 135, pages 565-570; July, 1976.

A HISTORY OF π. Petr Beckmann. The Golem Press, 1977.

PI AND THE AGM: A STUDY IN ANALYTIC NUMBER THEORY AND COMPUTATIONAL COMPLEXITY. Jonathan M. Borwein and Peter B. Borwein. John Wiley & Sons, Inc., 1986.

THE BUBONIC PLAGUE

THE BLACK DEATH. Philip Ziegler. Harper & Row, Inc., 1970.

LES HOMMES ET LA PESTE EN FRANCE ET DANS LES PAYS EUROPÉENS ET MÉDITERRANÉENS. Jean-Noël Biraben. Mouton et Cie., 1975.

THE BLACK DEATH IN THE MIDDLE EAST. Michael W. Dols. Princeton University Press, 1977.

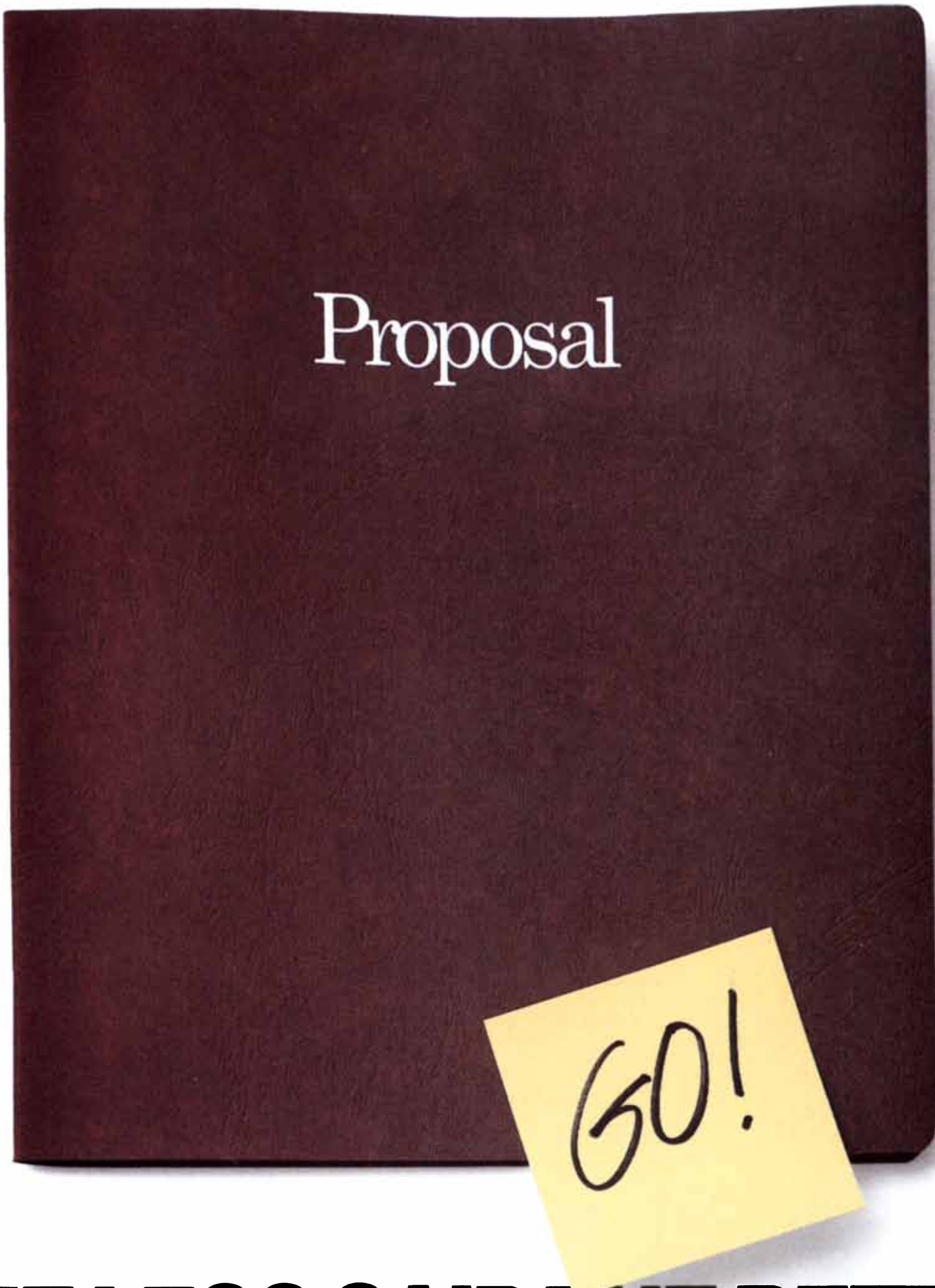
THE AMATEUR SCIENTIST

WATER WAVES: THE MATHEMATICAL THEORY WITH APPLICATIONS. J. J. Stoker. Interscience Publishers, 1957.

ELEMENTARY DERIVATION OF THE WAKE PATTERN OF A BOAT. Frank S. Crawford in *American Journal of Physics,* Vol. 52, No. 9, pages 782-785; September, 1984.

COMPUTER RECREATIONS

THE BLIND WATCHMAKER. Richard Dawkins. W. W. Norton & Company, 1987.



THE LESS SAID THE BETTER.

When People Count On You,
Count On Post-it™ Notes.
BRAND

"Post-it" is a trademark of 3M. ©3M 1987