SCIENTIFIC AMERICAN

AUGUST 1993
$3.95

Can particles move faster than light?

Putting chaos theory to work.

Cyberspace comes to the living room.

Nearby galaxy *displays its true colors when it is photographed using a 130-year-old technique.*

SCIENTIFIC AMERICAN

August 1993 Volume 269 Number 2

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 1993 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. and Canada 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111, or fax: (212) 355-0408.

THE COVER photograph shows the nearby spiral galaxy M83. The fine details and delicate hues seen here attest to the capabilities of modern color astrophotography (see "A Universe of Color," by David F. Malin, page 72). Bluish light in the galaxy's spiral arms emanates from fiercely hot, young stars. Yellow-brown lanes of dust and gas spawn star-forming regions, which glow pink where newborn stars have excited surrounding hydrogen atoms. A haze of elderly, yellowish stars envelops the galaxy's central regions.

THE ILLUSTRATIONS

Cover photograph by David F. Malin, Anglo-Australian Observatory

Page	Source
45	Sygma
46	Johnny Johnson
47–49	Jared Schneidman/JSD
53	Bettmann Archive
54–55	Patricia J. Wynne
56–60	Boris Starosta
63	Robert Becker/Custom Medical Stock
66–70	Dimitry Schidlovsky
72–77	David F. Malin
78	Pamela O. Lama, U.S. Naval Surface Warfare Center
79	Chris Usher/Black Star
80–81	Michael Goodman
82–83	Louis M. Pecora
84	Michael Goodman
86	Patricia J. Wynne
87	Steve Robinson/Natural History Photographic Agency
88–91	Patricia J. Wynne
92	Natural History Photographic Agency (*left*), Richard K. LaVal/Animals Animals (*right*)
93	Jason Goltz
94–95	Roger M. Macklis
96	Courtesy of *Journal of the American Medical Association*
97–99	Roger M. Macklis
100–101	John McGrail
102–103	George Retseck; adapted from information supplied by BroadBand Technologies and Time Warner, Inc.
104	Johnny Johnson
105	Merry Alpern
106	Johnny Johnson
107	Geoffrey Wheeler/Black Star (*left*), Microsoft Corporation (*right*)
120–122	Andrew Christie

SCIENTIFIC AMERICAN®

Established 1845

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRINTED IN U.S.A.

LETTERS TO THE EDITORS

Relative Alternatives

In “Black Holes and the Centrifugal Force Paradox” [SCIENTIFIC AMERICAN, March], Marek Artur Abramowicz proposed that close to a black hole the centrifugal force acting on orbiting objects would push them inward. A different interpretation of the phenomenon is also possible, however.

In general relativity, all energy, including kinetic energy, has weight. When an object moves faster, its weight increases. For an object orbiting sufficiently close to a black hole, the increase in weight (force directed toward the center) more than compensates for the increase in the centrifugal force (away from the center). Consequently, the net inward force increases with the speed.

Both interpretations of the physical phenomenon are equivalent, assuming that they are appropriately translated. One interpretation is probably simpler and more natural for some problems, and the other is for other problems. In physics, it is almost always helpful to have more than one way to look at the same things.

DON N. PAGE
Department of Physics
University of Alberta
Edmonton

Sound Reasoning

I was fascinated, as I am sure many readers were, by “Listening with Two Ears,” by Masakazu Konishi [SCIENTIFIC AMERICAN, April]. I wonder if the author can hint at the mechanism that enables me to tell if a sound source presenting simultaneous and equal stimuli to both of my ears is in front or in back of me—or indeed, where else it might be along the central sagittal plane?

CYRIL SANGER
Englewood, N. J.

Konishi’s article reminded me of how our technological advances during the past few decades were long anticipated by biological developments through evolution. The author did not comment on the owl’s practice of rotating his head in one plane. I had assumed that this motion was related to vision: it broadens the base of the owl’s triangulation for fixing distance. Now I am curious about its use for auditory distance sensing.

JOSEPH BURLOCK
Poquoson, Va.

Konishi replies:
Confusion about whether a sound emanates from in front or in back of a listener occurs when the localization cues are symmetrically distributed along the central sagittal plane. For the owl, the distribution of binaural cues is complex and asymmetric, which helps in pinpointing the sound source. Moreover, the ruff of feathers around the owl’s face makes its ear more sensitive to sound in the front of the head than in the back—much as the shape of the human ear helps us.

Owls turn their head because their eyes do not move. Barn owls can accurately localize the source of a short burst of noise that ceases before the head begins to move. That motion is therefore not essential for the owl to locate a sound in two dimensions. We do not know, however, whether head rotation contributes to the aural measurement of distance.

Dirac and the Arts

I liked the article about my late husband, "P.A.M. Dirac and the Beauty of Physics" [SCIENTIFIC AMERICAN, May]. I have a few misgivings, however, and I hope the authors, R. Corby Hovis and Helge Kragh, will not mind my correcting them.

Paul Dirac adored music. Even my knitting had to stop for complete silence when he was listening. He was also a great admirer of art. Not only did he like beautiful things in our home, he was also a tireless museum fan. He made me read *War and Peace,* and he read a great many books that I suggested to him. Theater, movies, ballet: we never missed a good performance, even if we had to go to London or from Princeton to New York.

Because Kragh took so much trouble over the biography that he wrote and this very informative article, I am more than sorry that he did not contact either our daughters or me.

MARGIT W. DIRAC
Tallahassee, Fla.

Hovis and Kragh reply:

We did not intend to exaggerate Dirac's scientific single-mindedness, and we hope that our concise account did not mislead readers.

Mrs. Dirac and others have noted that he enjoyed visiting museums and occasionally attending concerts, plays and movies. Yet nothing in his upbringing, education, writings or reported utterances suggests that he ever developed a real appreciation for the arts, and several anecdotes suggest a certain naïveté about literature and music. We are led to conclude that Dirac had only a nodding acquaintance with the arts and the humanities, unlike some of his great scientific contemporaries, such as Bohr, Heisenberg, Oppenheimer and Schrödinger.

The main focus of *Dirac: A Scientific Biography* and our article was Dirac's life in science. Kragh did write to him in 1981 to arrange an interview and similarly wrote to Mrs. Dirac in 1987, but neither request received a response.

Save That Trash!

I view with considerable alarm the idea of using plasma vitrification as a quick fix for our waste disposal dilemma by turning all our trash into a nearly useless mass of slag ["Garbage in, Gravel out," by W. Wayt Gibbs, "Science and Business," SCIENTIFIC AMERICAN, May]. There is an unrestrained predisposition for industrialized nations to select advanced technologies to solve the problems created by other advanced technologies. Our "waste" heaps are the end of the trail for our exhaustible resources.

As virgin resources become more depleted, our trash will become a major source of certain metals, plastics and billions of tons of biomass energy. With tax or other incentives to use recycled materials, mining our mountains of waste could rapidly become a big industry. Let us not turn our last resource stockpile into gravel.

KIRSTEN LLAMAS
Miami, Fla.

Because of the volume of mail, letters to the editor cannot be acknowledged. Letters selected for publication may be edited for length and clarity.

50 AND 100 YEARS AGO

AUGUST 1943

"In a combative, newly-published book, 'The Wright Brothers, a Biography Authorized by Orville Wright,' Fred C. Kelly demonstrates what is incontestably true—that it took the editor of *Scientific American* a long time to come to the point of believing that claims for the early Wright flights were truthful. Nearly three years elapsed between the Wrights' first powered flight and this magazine's full acknowledgement, in the number for December 15, 1906, of 'their epoch-making invention of the first successful flying machine.' In an age of publicity writers this slowness will be difficult to grasp. Let us go back. The Wrights flew and flew and flew on a field near Dayton, Ohio, in 1904 and 1905, in plain sight of a sightless world. They had plenty of troubles but worked up to five-minute flights, 18-minute flights, 25-minute flights, 38-minute flights, but it still wasn't news! The enterprising Dayton reporters obviously weren't so enterprising as our trusting editors believed. It was they, primarily, who kept the Wrights' big news in a vacuum. And Fred Kelly at the time—so says his own book—was a reporter dwelling only 11 miles from the Wrights' experiments!"

"Astonishing 30-day cures of long established hives cases resulted from oral administration of a drug which neutralizes histamine. The same drug relieved the histamine-sensitive patients of skin eruptions and acid stomach. Even rheumatoid arthritis and swelling of the legs and arms have been benefited. Dr. Louis E. Prickman, of the University of Minnesota, believes that antihistamine therapy offers great possibilities in the correction of food allergies."

"Now is revealed the part played by the American radio industry functioning in co-operation with the United States Navy and Army Signal Corps in the development of the revolutionary wartime science of detecting and ranging by radio. Basic research work was instituted by the Radio Corporation of America as early as 1932. During 1937, operating equipment was completed and tested, indicating the distance and position of reflecting objects, in much the same form as is now used in a large part of modern radar equipment. Westinghouse and RCA produced for the Signal Corps portions of its first radar apparatus, such as was in operation at Pearl Harbor, on December 7, 1941. It is a matter of record how radar warned of the approach of Japanese planes on that fateful morning, but the operator's report went unheeded. In September, 1940, it was radar that enabled the outnumbered Royal Air Force to turn back Hitler's previously invincible *Luftwaffe*."

SCIENTIFIC AMERICAN

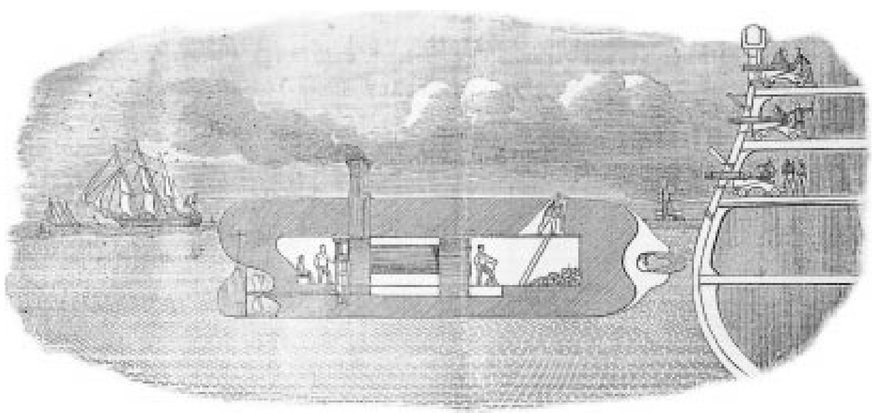
AUGUST 1893

"The Department of Agriculture has sent out circulars making inquiries over a wide extent of territory regarding the 'seventeen-year locusts,' which have made an appearance this year in eight States of the Union. The object of the department is to ascertain accurately the limits of the areas occupied by the insects. Strictly speaking, the insects are not locusts, but cicadae. Some years ago it was sought to introduce these insects as an article of diet; but the experiments in that direction did not promise success."

"Judging from the remains of extensive ancient works of irrigation, it is safe to say that the principal canals constructed and used by the ancient inhabitants of the Salado Valley in Arizona controlled the irrigation of at least 250,000 acres. Their canals are models for the modern farmer to imitate; yet they could have been dug in no conceivable manner save by the laborious process of hand excavation with stone or wooden implements."

"After three years' preparation the Polar expedition under Dr. Fridtjof Nansen has finally sailed from Christiania, Norway, for the North in the good ship *Fram* (*Advance*), the first vessel that has been especially designed and constructed for Arctic research. The vessel is a model of strength, but she is a trifle too small for the enormous amount of stores with which she has been loaded and which have brought her so down in the water that the ice sheathing has had to be heightened. Lighting will be electric or by means of lamps. The dynamo is worked either by steam, wind or hand power. A large windmill will be erected on deck, where there is also a winch which can be worked by four or more men, and, in order to give the hands exercise during the darkness, the latter will be daily resorted to in the winter months. The furnaces are constructed to burn petroleum, or even blubber, and under petroleum the vessel obtained on a trial trip the same speed as with coal."

"More than forty years ago, to wit, February 19, 1853, the SCIENTIFIC AMERICAN published illustrations of James Nasmyth's torpedo boat (*below*). Peculiar interest attaches to this submarine boat from the fact that a selection is soon to be made, by a board of examiners of the Navy Department, of a type of submarine vessel, for the construction of which Congress has appropriated $200,000. The principles of Mr. Nasmyth's floating mortar consist, in the first place, of a monster self-exploding shell, which is part and parcel of the vessel. The explosion of the shell is absorbed by the entire mass of the floating mortar."



Nasmyth's torpedo boat

Who Is Normal?

Is trying to "fix" a disability sometimes a mistake?

Four-year-old Jeremy Scharf is mischievous, outgoing—and profoundly deaf. Since this past March, however, when physicians at the Johns Hopkins University School of Medicine activated an electronic implant in his left ear, he has become an avid fan of birdsongs and music boxes. The implant takes over the functions of the boy's defective cochlea, the organ that sends signals to the auditory nerve. His mother, Roni, recalls that Jeremy recently complained about the noise she was making while emptying the dishwasher. "He told me to be quiet," she says. "It was wonderful."

Most onlookers might consider the availability of such devices an unalloyed blessing. Yet many people who are deaf or have other disabilities complain that attempts to devise medical "fixes" for their conditions are sometimes dangerously misguided. To Nancy Bloch, who is deaf and the executive director of the National Association of the Deaf (NAD), cochlear implants for children are so untried that they amount to "medical experimentation. As dirty as it sounds, that's exactly what it is."

Lee Kitchens, the president of the International Growth Foundation and a past president of the Little People of America, has similarly harsh words for surgical therapies designed to make dwarfs taller. "Instead of trying to modify the environment to fit the people, they're trying to modify the people," he says. "We think that's stupid."

At issue are questions about whether deafness, dwarfism and other disabilities should be regarded primarily as pathologies or as part of the normal spectrum of human variation. Medical opinions evolve over time. Homosexuality was once classified as a mental illness, but psychologists no longer call it one. Alcoholism was formerly a vice; now it is a disease. Accompanying those shifts were changes in attitudes about whether the conditions could—or should—be cured.

At the center of the current disputes are young children like Jeremy, whose parents make those decisions for them. The parents naturally want what is best for their kids—and understandably enough, that often means making the youngsters more like themselves. Parents worry, for example, that their deaf children will not hear approaching traffic or other warning sounds. "When Louis was very young, an angry dog bit him because he didn't hear it growling," says Judy Weiss of Bethesda, Md., whose son became one of the first children to receive an experimental cochlear implant 11 years ago. A larger parental concern is that deaf children will be shut out of social contacts and jobs if sign language, rather than English, is their native tongue. "We made the decision that we wanted Jeremy to be as much of a part of the hearing world as possible," Roni Scharf says.

JOHN TROHA *Black Star*

COCHLEAR IMPLANT gives Louis Weiss some hearing, but many deaf people argue that such devices are still dangerously experimental for young children.

"I don't think that even the most radical members of the deaf community would be able to make a very good case that deaf people are well integrated into society at large," comments Robert Shannon, director of research at the House Ear Institute in Los Angeles, where much of the early work on cochlear implants was done. "They aren't, and they cannot be, because most of our cultural interactions occur through spoken language."

Shannon believes the public should become more aware of the needs and talents of the deaf, but he also thinks it is important that people be free to make choices. "If there's a way we can overcome the hearing problems that these people have, why should we ignore it? If I had a vision problem, and somebody handed me a pair of glasses, I'd certainly wear them," he says.

Without question, some implant recipients have thrived. Jeremy's parents say his speech has been improving and that they do not intend to teach him sign language. Louis communicates with his family by talking and through a form of signing called cued speech; at school, he talks and lip-reads. According to his mother, Louis has always been mainstreamed in public schools, is an A student in the eighth grade and recently received an award for outstanding achievement in Spanish.

Not all those who have cochlear implants are so lucky. Even in the best cases, the implants cannot confer normal hearing. Both advocates and critics of the devices say only about 20 percent of the implant recipients hear well enough to understand most spoken sentences and to use the telephone easily. Perhaps an equal number derive virtually no benefit. The majority of the recipients fall somewhere in between: to varying degrees, the sounds they hear supplement their lipreading and environmental awareness. Generally the implants work best for those who lost their hearing after learning to speak; they are least effective for adults who have been deaf since early childhood.

Harlan Lane of Northeastern University, a hearing man, chairs the NAD's cochlear implant study group. In his opinion, the unreliability of the implants makes them risky because "there's a danger the child might end up in a no-man's-land." If the child's hearing and speech are poor, he or she will still be at a severe disadvantage in the hearing community. Moreover, without a knowledge of sign language, the child will also be an outsider among the deaf.

"The indisputable point is that the Food and Drug Administration did not consult any deaf people in its 1991 decision to authorize the implant for use in children," Lane says. "I think that's scandalous." He faults the dozen studies of the effectiveness of implants in children for methodological weaknesses and deplores the lack of research on the influence of the implants on children's psychological development. By ignoring those concerns, Lane argues, the medical establishment is treating "the deaf child as an ear with nothing attached." He believes that later this year the NAD may approach the FDA about reconsidering its authorization.

The deaf maintain that misconceptions about them are so pervasive that most hearing parents cannot make informed choices about the deaf way of life. "Many deaf people function in both worlds," Bloch says. Nearly all spend most of their time around hearing people, including ones in their own families. Many who have hearing impairments can still use the telephone to some extent. Keyboards and teletype displays attached to normal telephones, electronic mail and fax machines enable even the profoundly deaf to communicate by wire. Bloch, for instance, was interviewed for this story by telephone through a human interpreter and by fax.

"We consider ourselves more of a cultural group than a medical anomaly," Bloch explains, and as such, they are entitled to the respect due any ethnic, cultural or religious minority. The deaf have their own language, customs and history; unfortunately, their eloquence is lost on people who are illiterate in sign language. Because the real problem of the deaf is one of communication, Bloch contends, it should be solved by a social remedy, not a medical one.

Like the deaf, many Little People also sometimes find themselves at odds with parents. Campbell Howard, an endocrinologist in Kansas City, Mo., and the board president of the Human Growth Foundation, thinks adult Little People do not see a need to change the height of unusually short kids. Nevertheless, he adds, "a lot of parents come to me looking for something to make their children taller. They perceive a problem."

Kitchens says that what most disturbs people of short stature are "unnatural" attempts to make them taller. He is highly critical of limb-lengthening surgery, in which the long bones of dwarf children are repeatedly broken to stimulate their growth, because the procedure is painful, potentially harmful and only

But He'd Have to Leave the Cigars Behind

The U.S., as a member of the Pan-American Health Organization, frequently sends medical researchers to other countries to help investigate and manage disease outbreaks. Another member of the organization is Cuba. Although on frosty official terms with the U.S. and the subject of a trade embargo, Cuba is entitled to summon medical assistance.

That is what it did formally on April 5, as cases accumulated of an unknown illness characterized by impaired vision and loss of sensation. More than 40,000 cases have now been reported on the island, and although Cuban scientists had isolated a virus from some patients, they have been unable to prove it is the cause of the disease. Nutritional factors are suspected of playing a role, possibly in combination with a neurotoxin.

One of the U.S. scientists who went to Cuba to investigate was Paul W. Brown, a researcher at the National Institute of Neurological Disorders and Stroke and an expert on infectious diseases of the nervous system. Brown, who says he is uncertain about the cause of the strange illness, reports that U.S. scientists working in Cuba were surprised to be joined for two hours each evening during their discussions by Fidel Castro, Cuba's bearded and long-reigning revolutionary president.

Not only did Castro attend meetings, Brown says, he asked penetrating questions and frequently—and accurately—corrected scientists on their technical slips. "The man is amazing," Brown declares. "He has a mind like a steel trap. It's not hard to see why he's in charge." Brown and his colleagues returned to the U.S. with samples of spinal fluid and will try to duplicate the Cubans' isolation of a virus and make antibodies to the Cuban isolate. The results should make it possible to confirm or rule out the virus hypothesis. Brown doubts a virus is responsible but says he was impressed by the Cuban researchers' technical expertise: "We told Castro that if he wanted a new job we'd be pleased to have him at Bethesda as a colleague." Fidel declined the offer.

—Tim Beardsley

AGENCE FRANCE-PRESSE

PATIENT, one of thousands suffering from symptoms of a mysterious illness that affects sight and sensation, is tested in a Cuban clinic.

marginally effective. A small number of dwarfism cases are caused by a deficiency of growth hormone. Physicians can make such children grow with injections of a synthetic substitute. That therapy is unobjectionable, in Kitchens's view, because "it's much like people taking insulin because their pancreas is not working as it should. But in the case of a person who is three feet tall, a few inches isn't going to make much difference. The aggravation of the shots may be worse, from a psychological standpoint, than the gain you're going to get."

He believes "90 percent of short-statured people would say forget it" if given the chance to be average height. "I think I've led a fairly successful life." Kitchens, who is four feet tall, retired after 42 years as an engineer and teacher. He is now serving his third term as the mayor of Ransom Canyon, Tex. He quotes a joke made by a former president of the Little People of America: "You're never too short as long as you can reach the pedals of your Cadillac."

As genetic engineering and medical technology advance further, the opportunities to alter physical and mental characteristics will only increase. The decisions that will be made will undoubtedly be biased by social and cultural concerns—What is normal? What is desirable? But the availability of a procedure can subtly shape those attitudes. Although the use of growth hormone is sanctioned for boosting the height only of people with hormonal dwarfism, Howard reports that many parents put pressure on doctors to prescribe it for children who are just shorter than average. No one yet knows, he says, whether these hormonally normal children do get significantly taller or what the long-term side effects might be.

Kitchens, for one, feels there are advantages to being unusual. "If you are different, people remember you," he says. "You stand out in the crowd—if they can see you, that is."—*John Rennie*

Strange Matters

Can advanced accelerators initiate runaway reactions?

If you have trouble sleeping, you don't want to know about the physicist's worst nightmare: an atom smasher produces a new form of matter even more stable than everyday protons and neutrons, thereby triggering a cataclysmic, self-sustaining reaction that consumes the earth.

Although no serious scientists believe an atomic collision could ever lead to a global meltdown, they still want to be very, very sure it will never happen. Since the beginning of the nuclear age, researchers have met many times—usually behind closed doors—to discuss whether there was any chance that a proposed experiment might initiate a catastrophic event. Physicists rarely discuss the issue openly, fearing bad public relations, but recently some have given candid accounts of the secret meetings. "It's a real concern," observes Henry J. Crawford of the University of California at Berkeley. "Whenever scientists have started a new accelerator program, one of the first talks is always on this topic."

Indeed, one of the most astonishing debates of this subject was revealed by Subal Das Gupta and Gary D. Westfall in *Physics Today*. The story began some 30 years ago, when the Lawrence Berkeley Laboratory was planning to build a particle accelerator called the Bevalac. At the time, two theorists, Nobel laureate Tsung Dao Lee and the late Gian-Carlo Wick, raised the possibility that conditions of extreme energy and density could create a new phase of dense and stable nuclear matter. If this substance, known as Lee-Wick matter, existed and could be generated, the physicists feared, it would quickly accrete every atom around it—namely, the laboratory, California and the rest of the planet.

Researchers realized that the Bevalac had a shot at making Lee-Wick matter, and under no circumstances did they want to prove the theorists right during a test run of the machine. "We took the issue very seriously," comments Westfall, who was a member of the Bevalac's scientific staff at the time. "We appointed a blue-ribbon committee to make sure there was no chance it would happen."

The committee, which included Miklos Gyulassy, who is now at Columbia University, met several times. Together they concluded that the Bevalac had no chance of initiating a nuclear disaster. The physicists reasoned that nature had already performed the relevant experiment: the earth, moon and all celestial bodies are constantly bombarded with an extraordinary number of high-energy particles that are produced by stars. Some of the particles collide with atoms on the earth and create conditions that equal or surpass anything the Bevalac could do. Yet the planet was still reassuringly here. Nor had any such event destroyed the moon, which had been struck by countless high-energy particles for at least a few billion years.

In the 1970s the operation of the Bevalac and other accelerators confirmed that Lee-Wick matter did not exist. This happy state of affairs can be explained. When an atomic nucleus collides with another and is compressed into a volume about one fourth its normal value, it expands in about a thousandth of a billionth of a billionth of a second. Nuclear matter that has been compressed somewhat is simply not stable.

But what happens if nuclear matter is compressed to more extreme densities? If two nuclei collide at energies a bit beyond those that modern atom smashers can achieve, the nuclei should transform into so-called strange matter. The protons and neutrons of an atom are themselves made up of quarks, and when the quarks collide at high energy, they may yield a heavier particle: the strange quark. The consensus among theorists is that certain combinations of strange quarks with others are stable.

Strange matter should grow through the accretion of ordinary atoms. But not to worry. The droplet of matter should not get much larger than a few million strange particles, theorists think. All such particles should carry a relatively large quantity of positive charge that should ultimately cause the droplet to burst apart. "The basic idea is that at equilibrium the stuff has a net positive charge, and as a result it would turn its own reactions off," Crawford says.

So how can theorists be absolutely certain that an accelerator will never spawn a voracious clump of strange matter? The question was first posed seriously in 1983, when researchers were designing the Relativistic Heavy Ion Collider (RHIC). The collider, now under construction at Brookhaven National Laboratory, promises to be the world's most powerful smasher of heavy atoms and could quite possibly generate strange matter. Piet Hut of the Institute for Advanced Study in Princeton, N.J., put everyone's fears to rest. Applying the same logic his predecessors had used, Hut showed that innumerable cosmic particles collide with atoms on the earth and moon, creating conditions far more extreme than those of RHIC. Calculations similar to Hut's have been done "for all the accelerators that have been built so far," Crawford says, and therefore physicists know they are "not going to be walking in any dangerous territory."

Although there is no instrument yet built that could cause the earth to become a lump of strange matter, such transformations may occur in other celestial bodies. If a droplet of strange matter forms within a star made of dense neutral matter, it might initiate a chain reaction that would create a strange-matter star. Physicists say such events can occur only in the heavens. Let's hope they are right. —*Russell Ruthen*

Sound Science?

Researchers still sparring over effects of Exxon Valdez

The 1989 *Exxon Valdez* oil spill seems to have sullied more than the waters and wilderness of Prince William Sound. Because of lawsuits against the petroleum company, many studies about the condition of the ecosystem and wildlife were kept secret until this year. But even now, after the release of findings by Exxon and by the government, no consensus has been reached on what happened to the sound after the 10.8-million-gallon spill and whether it is or is not recovering.

Frustrated researchers say spin has subsumed science. "I find it very disturbing," comments Robert B. Spies, chief scientist for the trustees, a group of federal and state representatives who have overseen damage assessment studies. Investigators "come to opposite conclusions, and then the public asks, 'What good is science if the answer depends on where you are getting your money?'"

Scientists for all parties initially assumed they would share data and then make their own interpretations. Once lawsuits were initiated, however, Exxon and the government trustees banned the release of any information. Open discussion, debate and peer review were suspended. Many scientists continue to worry that the opportunity to learn from the spill was squandered because lawyers shaped the choice of studies. Some decisions about cleanup were also made without relevant data on, for instance, fish populations.

After the 1991 settlement—in which Exxon agreed to pay $1.1 billion to the state and federal governments—the long-awaited data promised to surface. Most of them finally did this past February in Anchorage. Researchers for the trustees presented their findings, cataloguing extensive and, in some cases, ongoing damage to many species of birds, fish and mammals. Although scientists from Exxon were invited to present their work at the same time, they declined.

In late April, Exxon responded with its own meeting in Atlanta. There experts offered a diametrically different view of the sound. They reported that the area was in fact entirely recovered and that contamination had been extensive before the spill. That assertion was based on studies of hydrocarbon fingerprinting, a means of characterizing the source of the oil. Exxon chemists say that before the spill the sound was already polluted by diesel oil and oil from a natural seep. They maintain that federal scientists mistakenly identified oil from these sources as *Exxon Valdez* oil.

"Exxon seems to want to make the claim that there was significant widespread contamination prior to the spill. We don't agree," says Jeffrey W. Short, a chemist at the National Oceanic and Atmospheric Administration (NOAA). "We did a four-year baseline study when the [shipping] terminal opened in 1977. In the course of that study of intertidal sediment, we didn't see any evidence of seep oil and precious little of diesel oil."

Exxon's conclusions about the intertidal region also differ from those of NOAA. Jonathan P. Houghton, a marine biologist who has studied the sound for both Exxon and NOAA, has reported that washing the beaches with hot water was often detrimental and that recovery of the flora and fauna has been slow in some places because of the cleanup. Using an alternative methodology and a different definition of an oiled beach, Exxon scientists reached another conclusion. "We feel, in general, the sound has essentially recovered," comments Alan Maki, chief scientist for Exxon.

It is unlikely that anything more conclusive than a contest of point-counterpoint can emerge from a comparison of the studies for now. The trustees' findings have been peer-reviewed and are currently available to the public. But the company itself has generally released only summaries of its data, and as of yet none of them has been vetted. Exxon faces some $2.6 billion in additional lawsuits brought by Native Americans and fishermen and is not expected to fully reveal its data until after hearings in 1994. In addition, a provision called the "reopener clause" in the 1991 settlement permits the trustees to sue for more money if they discover further damage to the sound.

Many researchers say they cannot wait to get their hands on the data base and sort through the varied interpretations. "I will do it on my own time," Spies exclaims. "I want to see what the scientific basis of all this is." He adds that his concern extends to both groups of researchers. A study by the trustees of abnormality in juvenile herring, for example, concluded that the population at large was affected. "But the data don't really support that yet. It is in a gray area," Spies cautions. On the other hand, given the same data, "Exxon scientists extrapolate that there wasn't an effect."

Despite concerns that have emerged about conducting science during a legal battle, another oil spill would probably give rise to a similar situation. Some experts have advocated establishing independent commissions in such cases. But the precedents are not entirely reassuring, notes Charles H. Peterson, professor of marine science at the University of North Carolina at Chapel Hill. In the 1970s, for instance, environmental organizations sought to block the expansion of the San Onofre nuclear power plant in California.

The utility and the activists agreed to pool funds to study the impact on the environment and to create a review committee—made up of members from industry, conservation groups and academia. Both sides agreed to accept the results. "The concept was good," Peterson says. But after 14 years and $50 million, "the company went off and did its own studies, so it all went to court anyway." —*Marguerite Holloway*

J. L. ATLAN Sygma

CRUDE OIL was washed from beaches after the* Exxon Valdez *spill. Exxon says recovery of the area is complete; government scientists say it is not.

STEVE STARR SABA

HIROO KANAMORI of the California Institute of Technology is working to outrace seismic waves to get earthquake data out to critical industries and rescue workers.

Fast Moves

Instant earthquake analysis may beat the waves

The ground must still tremble before seismologists can collect data on earthquakes, but the speed at which they can analyze the abrupt shifts of faults is surpassing the swiftness of the shock waves. Indeed, it is now possible to provide a crucial warning of a few seconds for some aftershocks. And while attempts to predict earthquakes have met with mixed success, researchers may soon be able to anticipate the effects of an initial tremor, enabling railroads to stop or slow their trains and permit elevator systems to halt at the nearest floor. "It is a tremendous development," says Hiroo Kanamori of the California Institute of Technology, which helps to operate one of the leading earthquake research centers.

Such warnings are becoming possible now that seismic analysis can be completed practically in real time. Seismic stations around the globe are linked by telephone lines, which gives them nearly instantaneous access to readings recorded after an earthquake. Seismologists transmit their analyses of the tremor—its location, depth, magnitude and the orientation of the fault, for instance—to their colleagues through the Internet with equal rapidity. Moreover, researchers have developed streamlined methods for picking the most relevant information out of the seismic data. "Analysis that would have taken many months 10 years ago now happens in hours or less," says Thorne Lay of the University of California at Santa Cruz.

Several groups across the U.S. engage in what Lay describes as "friendly competition" to see who can complete the work the fastest. In the case of the Landers earthquake, which struck southern California in June 1992, Lay's group managed to derive the complete geometry of the fault in just a few hours, "without ever leaving the office."

Taking that approach a step further, Göran Ekström and his collaborators at Harvard University are developing a fully automated system in which a computer collects information from the worldwide network of seismometers, processes it and relays the key attributes of each quake to the scientific community. "Seismologists can stay home and watch TV," Lay jokes. "But seriously, automation eases the tedium and allows us to use our time more creatively." Ekström is eager to eliminate "human intervention" so that he can disseminate predictions of tsunamis, which require five to 20 hours to travel across the Pacific Ocean. He hopes to be able to predict the height and location of these "tidal" waves only one to two hours after the initiating earthquake, well before the wall of water reaches the shore.

Improvements in the technology of seismic detectors are also assuring that the fast analysis of earthquakes is more accurate than ever before. The latest devices can record very strong ground motions that ran off the scale of previous instruments. "Now we don't lose data," Kanamori says. Being able to monitor the most intense part of the quakes "reveals aspects of rupture never before seen," Lay adds. These data are especially valuable because they capture details of the very early stages of disruption occurring along a fault.

Fast analysis of earthquakes permits researchers to properly deploy portable instruments immediately after the shock. Quick response is essential for learning more about the ground motions that follow as the earth's crust adjusts to its new stresses, Kanamori notes. Portable seismometers can also give short-term (about 20 seconds) warnings of aftershocks—enough time to help protect rescue workers sifting through unsteady rubble. One such temporary seismic network aided searches through the remains of the Oakland viaduct after California's severe Loma Prieta earthquake of 1989, according to Thomas Henyey of the University of Southern California, who is also the executive director of the Southern California Earthquake Center.

Much of the center's work focuses on learning more about "the potential damage scenario" after an earthquake, Henyey explains. Producing reliable estimates of the locations of the most severe ground motion is not a simple task because it requires knowing a great deal about local geology and about the nature of the earthquake source. Such estimates are extremely valuable, however, because they can guide rescue workers to the hardest-hit areas, help utilities and railroad companies target their repair efforts and direct engineers to the most heavily damaged buildings.

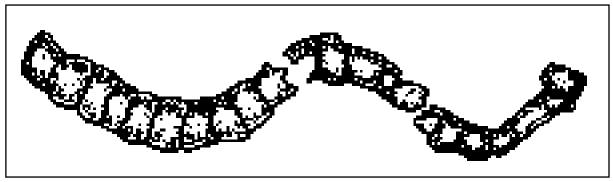
To that end, Kanamori, Egill Hauksson of Caltech and Thomas Heaton of the U.S. Geological Survey have spearheaded a new project, the Caltech-USGS Broadcast of Earthquakes, or CUBE. Kanamori describes CUBE as an automated paging system that sends out earthquake data to those who need it most: emergency services and the industries that supply water, power, telephone service and rail links. Through CUBE, earthquake researchers meet with representatives from local industry to discuss their needs. "We have been talking about this idea for years, about reaching beyond academia," Kanamori explains. "It is very important to involve people in the real world."

Because of its mission, CUBE faces the daunting challenge of issuing earthquake data that are not only timely but also exceedingly reliable. Fast-analysis systems like the one at Harvard need not worry about small errors, because "they're not intended to support emergency operations," Kanamori says. "CUBE cannot afford to make mistakes." Ordinary telephone links are too slow and unreliable during an earthquake, so CUBE uses dedicated telephone lines and microwave and radio links.

Of course, CUBE can be only as useful as the information it distributes. Present techniques require at least a few minutes to process data from an earthquake, "after everyone knows the shaking has started," Kanamori notes dryly. The goal is to "beat the seismic waves" by cutting the processing time to less than a minute, so that CUBE could become a bona fide early-warning system.

Right now CUBE concerns itself only with southern California, but Kanamori foresees that a coordinated network of similar earthquake-broadcast systems will soon cover all of California and eventually all of the U.S. Not surprisingly, companies in other countries are also pursuing the benefits of rapid earthquake analysis. In Japan, for instance, even insurance companies are getting in on the act as they cast a baleful eye on plans for ultra-tall skyscrapers that may be built on the seismically squirmy islands. *—Corey S. Powell*

Off to an Early Start



Determining when and how life began is maddeningly difficult. Erosion, volcanism and the ceaseless churning of tectonic plates have obliterated virtually all traces of the earth's early history. Only by analyzing meteors, which presumably coalesced at the same time as the earth, have investigators determined that the age of the planet itself is roughly 4.5 billion years.

This situation makes a recent report in *Science* by J. William Schopf of the University of California at Los Angeles, an authority on early life, all the more remarkable. Schopf presents fossils suggesting that microorganisms not only existed but had achieved a significant degree of complexity at least 3.465 billion years ago. The finding narrows the window of opportunity during which mere matter could become animate. It also lends support to a surprising scenario involving the formation of the modern atmosphere and of vast deposits of iron ore laid down billions of years ago.

Schopf's evidence consists of minute, filamentous impressions left by microbes in sedimentary rock from northwestern Australia. The age of the rocks was established by measuring the degree of radioactive decay in their constituent elements. "Life was flourishing back then," Schopf says.

The fossil microbes, which measure one to 20 microns wide and up to 90 microns long, were linked together like beads on a string [*see illustrations above*]. Based on variations in the size and shape of the individual cells—and particularly the cells capping the filaments—Schopf has identified at least 11 separate species. By comparing the fossils with modern prokaryotic organisms, he has concluded that a majority were probably cyanobacteria. Also called blue-green algae, they convert sunlight into energy through photosynthesis and excrete oxygen in the process.

Various workers, including Schopf himself, have previously reported finding fossils of individual microorganisms and of dense microbial colonies, known as stromalolites, more than three billion years old. Skeptics worried that the alleged fossils may have been improperly dated or even created by nonbiological processes. Schopf thinks his new results should put these doubts to rest. "This is real firm," he remarks. "It is the sort of thing that can get into the textbooks and stay there."

Schopf maintains that his data still allow plenty of time for rudimentary life-forms to develop. By studying craters on the moon, geologists have determined that the earth was bombarded by asteroids for hundreds of millions of years. Investigators believe these impacts may have rendered the earth uninhabitable until at least 3.9 million years ago.

Some reports have even indicated that given the time required for mere matter to assemble itself into life, the first organisms must have arrived on the earth from elsewhere—that is, from outer space. Schopf rejects this theory. "Four hundred million years is still 400 million years," he says. After all, he notes, the past 400 million years encompasses virtually the entire history of vertebrates, including amphibians, dinosaurs, mammals and *Homo sapiens*.

One person delighted by Schopf's results is Kenneth M. Towe, a paleobiologist at the Smithsonian Institution. For years, Towe has argued—based primarily on geologic data—that cyanobacteria caused oxygen to build up in the atmosphere not just two billion years ago, as conventional wisdom has it, but at least 3.5 billion years ago. Towe also suggests that the oxygen in the atmosphere combined with iron in the oceans to form iron oxide, or rust, that eventually settled on the seafloor. This process created iron ore deposits, called banded iron formations, that are found throughout the world.

But another recent paper supports a different scenario. A group led by Friedrich Widdel of the Max Planck Institute for Marine Biology in Bremen has discovered a type of bacteria that employs an unusual type of photosynthesis, which generates iron oxides rather than oxygen as a waste product. Widdel's group speculates in *Nature* that these organisms could have formed the banded iron formations without injecting oxygen into the atmosphere.

On the other hand, the iron deposits may have formed without any help from bacteria, according to James F. Kasting, a paleogeologist at Pennsylvania State University. Ultraviolet radiation, Kasting explains, could have knocked oxygen molecules free from water molecules; the oxygen would then have immediately reacted with iron in the oceans to form iron oxides and thus rust-laden sediments.

Unfortunately, Kasting observes, there are not enough data to prove—or rule out—any of these theories. When it comes to the origin and early evolution of life, some mysteries seem as intractable as ever. *—John Horgan*

Culture Clash

Is mathematics becoming too much like physics?

In this century, physicists have followed mathematicians' lead rather than vice versa. Albert Einstein fashioned his theory of relativity out of an exotic non-Euclidean geometry devised by Georg Riemann more than a century ago. Inventors of modern quantum theory exploited a theory of groups invented even earlier by Evariste Galois. The past decade or so has seen a remarkable reversal of this flow of ideas, as advances in physics—particularly in an esoteric field called superstring theory—have begun to inspire work in pure mathematics. The trend has been welcomed by some mathematicians, but others fear that their discipline is in danger of losing its moorings as it adopts the more speculative style of physics, in which mathematical proof has never been paramount.

These concerns have been aired in a controversial article in the July issue of the *Bulletin of the American Mathematical Society,* a journal received by all 30,000 members of the AMS. "Is speculative mathematics dangerous?" ask Arthur Jaffe of Harvard University and Frank S. Quinn of the Virginia Polytechnic Institute. "Recent interactions between physics and mathematics pose the question with some force: traditional mathematical norms discourage speculation; but it is the fabric of theoretical physics. In practice there can be benefits, but there can also be unpleasant and destructive consequences."

While some mathematicians applaud Jaffe and Quinn's complaint, others deplore it—and even question the decision of the *Bulletin* to publish it. The journal "usually does not publish this kind of thing," admits Richard S. Palais, a mathematician at Brandeis University and an editor of the publication. But he believes the issues raised by Jaffe and Quinn are too important to ignore. Mathematicians have been discussing the issues "on a gossip level" for years, Palais says. "It is a real service they've done by putting this down carefully into words."

In their article, which is titled "Theoretical Mathematics: Toward a Cultural Synthesis of Mathematics and Theoretical Physics," Jaffe and Quinn acknowledge that both mathematicians and physicists begin by posing conjectures about various phenomena. But in mathematics, progress requires that conjectures then be proved—or disproved—through a set of rigorously logical deductions. Proofs serve the same role in mathematics that experiments do in physics, they suggest, and history has shown which method is more reliable. After all, what theory of physics has had the shelf life of, say, the Pythagorean theorem? "Our literature is very long-lived," Quinn remarks.

History has also shown the dangers of too speculative a style. Early in this century, Jaffe and Quinn recall, the so-called Italian school of algebraic geometry "collapsed after a decade of brilliant speculation" when it became apparent that its fundamental assumptions had never been properly proved. Later mathematicians, unsure of the field's foundations, avoided it.

Jaffe and Quinn fear that the growing influence of superstring theory on mathematics may cause this pattern to recur. Developed in the early 1980s, superstring theory holds that all the particles and forces of nature—including gravity, electromagnetism and the nuclear forces—arise from the tremblings of infinitesimal loops of energy called superstrings. The theory lacks empirical evidence, and Edward Witten of the Institute for Advanced Study in Princeton, N.J., a preeminent superstring theorist, has proclaimed that progress in the field is most likely to come about by unearthing what he calls the theory's "core geometrical ideas."

In his quest for these ideas, Witten has ventured more and more deeply into mathematics. Borrowing from his own knowledge of quantum field theory, he has posed conjectures in knot theory and topology that have inspired a whole new industry among mathematicians. In 1990 Witten won the Fields Medal, the most prestigious award in mathematics, for his work.

Some mathematicians grumbled that the medal should be reserved for those who had actually devised proofs and not merely posed conjectures, no matter how interesting. Jaffe insists he thinks "extremely highly of Ed Witten" and believes he deserved the medal. Yet Jaffe worries that the mathematics community might be sending the wrong message to young mathematicians by implying that speculative ideas are more important than "the hard work of proofs."

He and Quinn recommend a set of "prescriptions" to minimize damage caused by "basing mathematics on intuitive reasoning, without proof." Their first and most important recommendation is that conjectural work should be clearly distinguished from rigorous proofs. "Referees and editors should enforce this distinction," they say, "and it should be included in the education of students." They also argue that "a major share of credit should be reserved for the rigorous work that validates" mathematical conjectures.

Jaffe and Quinn have managed to provoke some prominent mathematicians. One is William P. Thurston of the University of California at Berkeley, whom Jaffe and Quinn accuse of having provided an insufficient proof for a major topological conjecture a decade ago. As a result, they complain, "a grand insight delivered with beautiful but insufficient hints" became "a roadblock rather than an inspiration." Thurston replies that he has proved one special case of the conjecture and that "there is a huge array of evidence that the conjecture is true in general." He adds, "What's important is that there is a body of mathematicians who are producing interesting mathematics based on this."

Indeed, Thurston contends that the distinction between speculative and rigorous mathematics is not nearly as clear-cut as Jaffe and Quinn imply. "The idea that mathematics reduces to a set of formal proofs is itself a shaky idea," he says. "In practice, mathematicians prove theorems in a social context. I think mathematics that is highly formalized is more likely to be wrong" than mathematics that is intuitive.

Isadore M. Singer, a mathematician at the Massachusetts Institute of Technology, thinks mathematicians are much more capable of distinguishing speculation from rigorous proof than Jaffe and Quinn suggest. He also dismisses their concern that those who prove conjectures put forth by others may receive too little credit. He notes that one conjecture in topology originally posed by Witten was recently proved by a young Russian mathematician named Maxim Kontsevich. "Now Kontsevich has become famous," Singer says.

As for Witten, he feels that Jaffe and Quinn do not sufficiently appreciate the power and depth of superstring theory. "Their description of the status of string theory is very narrow," Witten comments. "They also don't attempt to convey the importance of the ideas of physics in mathematics."

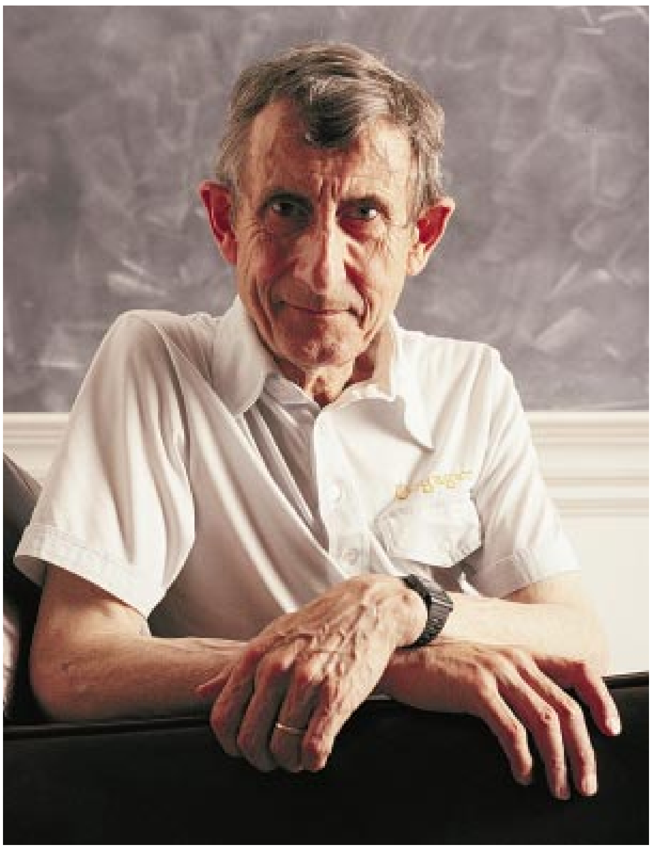
Jaffe and Quinn have their supporters. Richard M. Schoen, a mathematician at the Institute for Advanced Study, agrees with them that uncertainty over whether a given theorem is actually proved can create a "dead field." Stephen Smale of Berkeley calls Jaffe and Quinn "courageous" for their insistence on the value of old-fashioned rigor, although he disagrees with them about the need for specific rules to uphold standards. "The important thing is for people to become more conscious of these issues," Smale observes, "and not to have a lot of rules." —*John Horgan*

Perpendicular to the Mainstream

Some particle physicists think of Freeman J. Dyson as a rather tragic figure. In the early 1950s, they recall, the British-born physicist was at the very center of the field, striving with Richard P. Feynman and other titans to forge a quantum theory of electromagnetism. Dyson, some say, deserved a Nobel Prize for his efforts—or at least more credit. They also suggest that disappointment and, perhaps, a contrary streak later drove him toward pursuits unworthy of his powers. “Freeman always has to be perpendicular to the mainstream,” one theorist says.

When I mention this assessment to Dyson in his office at the Institute for Advanced Study in Princeton, N.J., his base for 40 years, he gives a tight-lipped smile. He then responds, as he is wont to do, with an anecdote. The British physicist Lawrence Bragg, he notes, was “a sort of role model.” After Bragg became the director of the University of Cambridge’s Cavendish Laboratory in 1938, he steered it away from nuclear physics, on which its mighty reputation rested, and into new territory. “Everybody thought Bragg was destroying the Cavendish by getting out of the mainstream,” Dyson says. “But of course it was a wonderful decision, because he brought in molecular biology and radio astronomy. Those are the two things that made Cambridge famous over the next 30 years or so.”

Dyson has spent his career swerving toward unknown lands. He has veered from mathematics, his focus in college, to particle physics and from there to solid-state physics, nuclear engineering and climate studies, among other fields. Dyson is probably best known now for his books, the first of which, the memoir *Disturbing the Universe,* was published in 1979. His writings celebrate diversity, both in their subject matter—which ranges from the origin of life to the long-term prospects for intelligence in the cosmos—and as a principle in itself. “The principle of maximum diversity,” he wrote in his 1988 book *Infinite in All Directions,* ensures that “when things are dull, something new turns up to challenge us and stop us from settling into a rut.”



ROBERT PROCHNOW

“God is in the details” is one of Dyson’s favorite sayings.

In person, too, Dyson keeps thwarting expectations. At 69, he is slight, all sinew and veins, with a cutlass of a nose and deep-set, watchful eyes. He resembles a raptor, a gentle one. His demeanor is cool, reserved—until he laughs. Then he snorts repeatedly, shoulders heaving, like a 12-year-old schoolboy hearing a dirty joke. It is a subversive laugh, that of a man who envisions space as a haven for “religious fanatics,” “recalcitrant teenagers” and other misfits, and who insists that science at its best must be “a rebellion against authority.”

Dyson’s father, a musician, headed the Royal College of Music, and his mother held a law degree (but never practiced). Their boy displayed his scientific and literary talents early. He calculated the number of atoms in the sun at six and began writing a science-fiction novel at eight. Entering Cambridge in 1941, he quickly developed a reputation as one of England’s most promising mathematicians.

A pacifist before World War II, Dyson decided that such a position was politically untenable after Germany overran France. During the war, he spent two years working for the Royal Air Force, seeking ways to reduce casualties among its bomber crews. “I learned everything I know about war in those two years,” he says. “The whole bureaucracy was designed so that the commander in chief would hear what he wanted to hear, which is equally true now, of course.”

Resuming his studies at Cambridge, Dyson became increasingly attracted to theoretical physics. In a history of quantum electrodynamics that is to be published this fall, Silvan S. Schweber of Brandeis University recounts how Dyson revealed his decision to an acquaintance at Cambridge. As they strolled through the campus, the colleague said, “Theoretical physics is in such a mess, I have decided to switch to pure mathematics.” “That’s curious,” Dyson replied. “I have decided to switch to theoretical physics for precisely the same reason.”

In 1947 Dyson traveled to Cornell University to study under the great physicist Hans A. Bethe. There he befriended Feynman, whom Dyson once described as “half genius and half buffoon.” Feynman had invented an idiosyncratic method for describing electromagnetic interactions. Meanwhile Julian Schwinger of Harvard University had proposed a seemingly different theory. In a series of brilliant papers, Dyson demonstrated that the two theories were mathematically equivalent. He went on to champion a more lucid version of Feynman's method.

Some physicists have argued that Dyson's contributions were as crucial as those of Schwinger and perhaps even Feynman. Dyson demurs. "I was a clarifier, not an inventor," he says. He insists he has no regrets about not receiving the Nobel Prize, which Schwinger and Feynman shared in 1965 with the Japanese physicist Sin-Itiro Tomonaga (who had derived the theory independently). "I suppose I'm lucky I never succumbed to this Nobel disease that many of my friends seem to have suffered from," he says. "It certainly never played a role in motivating me."

By the mid-1950s Dyson had moved from Cornell to the Institute for Advanced Study, and he decided it was time to change fields. "Particle physics had become the fashionable mainstream, and there were all these piles of preprints coming in every day. I felt just adding one more to the pile wasn't really worthwhile." Once he abandoned the search for a unified theory of physics, Dyson never returned. "I'm not really interested in the big picture," he says. "'God is in the details'—that's one of my favorite quotes."

Switching to solid-state physics, he became absorbed in spin waves, the oscillations of atomic spins in a ferromagnet. Saying "I'd like to brag just a little bit," Dyson points out that his 1956 paper on spin waves has been cited at least 675 times and was recently chosen as a "citation classic" by the journal *Current Contents.* "From the point of view of the community at large," he remarks wryly, "you might say spin waves is really the most important thing I did."

In 1956 Dyson also acquired "the delightful hobby" of engineering when he became a consultant to General Atomics, a company dedicated to the peaceful uses of nuclear energy. He helped to invent a miniature reactor that generates radioactive isotopes for research and medical applications. "We designed the thing within two months, built it and sold it in two years." Dyson received his only patent for the reactor, which is still in use throughout the world.

Dyson indulged a long-standing obsession with space exploration by helping General Atomics design a nuclear-powered spaceship called *Orion.* That project ended unsuccessfully in 1965, and Dyson now thinks nuclear rockets "are probably not much good." He remains interested in alternative approaches to space propulsion. One method he favors calls for using a powerful laser (based on a mountaintop, perhaps) to vaporize water or other propellants in spaceships and accelerate them skyward. "It has this fatal flaw," he admits. "It's only cheap if you have a high volume of traffic."

Unfortunately, in the late 1970s the National Aeronautics and Space Administration cut off almost all funding for research on space propulsion technologies. Instead NASA funneled its resources into one large, expensive machine, the shuttle, violating all the principles Dyson holds dear. "It was just absolute stupidity of the worst sort," Dyson fumes. He notes that "apart from the fact that seven people were killed," he welcomed the destruction of the *Challenger*, because he thought it would lead NASA to abandon the shuttle once and for all. "It hasn't been that easy, but I still think we will get rid of it."

Dyson foresees no limits—cognitive, spatial or temporal—to the growth of intelligence.

The best way to revitalize NASA, Dyson contends, is to "cut it off from the octopus in Washington" and dismantle it, much as AT&T was dismantled. He remains confident that one way or another humanity—if not the U.S.—will fulfill its destiny in space. "The rest of the world is doing very well, thank you, particularly France and Japan."

Moreover, the greatest threat to civilization is receding. For several decades, Dyson has been a member of Jason, a group of scientists that advises the U.S. on national security issues. Just five years ago he and other Jason members spent the day with a nuclear-bomber crew at a Strategic Air Command base. "I had my hand on the red switch that arms the bombs," he says. Now the planes have been taken off alert and the bombs placed in storage. "That's enormous progress. Of course, there are still huge problems in destroying weapons."

Dyson's sense of whimsy and romance assert themselves as he peers farther into the future. He is not one of those who believes humans will soon be superseded by robots or computers. "Biological machinery is in so many ways more flexible than computer hardware," he says. In fact, he has speculated that one day genetic engineers may be able to "grow" spacecraft "about as big as a chicken and about as smart," which can flit on sunlight-powered wings through the solar system and beyond, acting as our scouts. Dyson calls them "astrochickens." He has also proposed that very advanced civilizations, perhaps concerned about dwindling energy supplies, could capture the radiation of stars by constructing shells around them.

In 1979 Dyson revealed the depths of his optimism in one of the more exotic papers ever published in *Reviews of Modern Physics.* Dyson had been piqued by a statement made by the physicist Steven Weinberg in his book *The First Three Minutes:* "The more the universe seems comprehensible, the more it also seems pointless." No universe with intelligence is pointless, Dyson retorted. He then sought to show that intelligence could persist for eternity—perhaps in the form of a cloud of charged particles—through shrewd conservation of energy. "No matter how far we go into the future, there will always be new things happening, new information coming in, new worlds to explore, a constantly expanding domain of life, consciousness and memory," Dyson proclaimed.

Dyson is "open-minded" about the possibility that in our cosmic journeys we will bump into aliens. "The closest I will come to an alien intelligence," he says, is an autistic woman he has known since she was a child. When she was 10, another autistic child sent her a letter consisting entirely of numbers. After scanning the list for a moment, the girl shouted, "Mistake! Mistake!" It turned out that all the numbers were prime but one—the mistake. "Although she comes from this totally different universe, mathematics is something she can share with us," Dyson explains. "Maybe that will be true of aliens, too."

Next spring the Institute for Advanced Study plans to hold a festival to honor the official retirement of its most veteran active member. Dyson may harbor some ambivalence about having been there so long. When I remark that the institute is the ultimate ivory tower, he nods solemnly. "I try to get out as often as I can to remind myself there's a real world," he says, and snickers.

For several years, in fact, Dyson has been traveling to colleges around the country talking to students. He prefers undergraduates, since "graduate students are narrowly focused on some rather unimportant problems, and they just don't seem to enjoy life very much." Dyson also likes small colleges in obscure locales. Last fall he visited the Vermillion campus of the University of South Dakota, and there he heard an "absolutely superb" concert of 16th-century music. "Someone who had met me at the airport said, 'Oh, lucky your plane was early, you're just in time for the concert.'" Dyson's seamed face brightens at the memory. Oh, the wonders one encounters, his expression seems to say, once one ventures outside the mainstream. *—John Horgan*

SCIENTIFIC AMERICAN

Eliminating Nuclear Warheads

More than 50,000 nuclear weapons may be decommissioned during the next 10 years. Their disposal requires both technical and political innovations

by Frank von Hippel, Marvin Miller, Harold Feiveson, Anatoli Diakov and Frans Berkhout

The U.S. and the former Soviet Union are making deep cuts in their cold war arsenals. In the long run, the elimination of tens of thousands of surplus nuclear weapons will greatly reduce the threat of nuclear war. In the short term, however, chaotic conditions in the former Soviet Union pose a danger that weapons or materials derived from them may find their way to renegade states or terrorist groups.

About 35,000 nuclear warheads are scattered across the territory of four of the nations that were born when the Soviet Union disintegrated late in 1991: Russia, Ukraine, Kazakhstan and Belarus. Political struggle persists within Russia, which inherited the largest part of the arsenal, as does friction between Russia and Ukraine, which inherited the second largest part.

Some progress has been made in securing the surplus nuclear warheads. All Soviet tactical warheads deployed in the 14 non-Russian republics and most of those deployed in Russia have reportedly been withdrawn to storage sites within Russia, significantly reducing the risk of unauthorized use or theft.

In addition, the START I and START II agreements—signed in July 1991 and January 1993, respectively—would have the former Soviet Union and the U.S. reduce their strategic arsenals from roughly 10,000 warheads apiece today to less than 3,500 each by the year 2003. Under START I, Ukraine, Kazakhstan and Belarus have agreed to remove the approximately 3,000 strategic nuclear warheads that remain in their territories to Russia for dismantling and to join the Nonproliferation Treaty as nonnuclear weapons states. Belarus has ratified both treaties, but Kazakhstan has ratified only START I, and Ukraine has ratified neither. Moreover, Russian hardliners may oppose ratification of START II because it would eliminate multiple-warhead land-based missiles, the heart of the Russian strategic arsenal, while leaving U.S. submarine and bomber forces essentially intact.

Even if all these treaties are ratified, the problem of implementing them will remain. The unsettled political situation in Russia has put its nuclear complex under extraordinary stress. In December 1992 the head of the Russian nuclear-fuel reprocessing facility outside Chelyabinsk, where more than 25 tons of separated plutonium is stored, complained that his workers had not been paid in more than two months. Scientists in Russia's nuclear-weapons design laboratories were told earlier that year to plant potatoes if they wanted to be sure to have food for their families.

Transporting tens of thousands of decommissioned nuclear weapons to storage locations, dismantling them and disposing securely of their uranium and plutonium will be a daunting task, especially under the current circumstances. There are no confirmed reports that Soviet warheads or materials have been diverted, but it is imperative that arrangements be agreed on that will allow monitoring and assistance from the West.

Comparable security concerns do not exist today in the U.S. warhead elimination process. Nevertheless, political considerations require that monitoring be done on a reciprocal basis. Indeed, the U.S. Senate recognized this fact when it ratified START I in October 1992 and instructed the president to seek agreement on reciprocal inspections and other means to monitor the numbers of nuclear weapons in the stockpiles of the U.S. and the former Soviet Union. The Russian government has indicated that it would accept such reciprocal monitoring, but thus far the U.S. has focused on trying to negotiate unilateral U.S. monitoring of aspects of Russian warhead elimination.

This policy should be reconsidered. What progress has been made to date has been a result of U.S. willingness to make reciprocal concessions, such as the matching "unilateral" initiatives, an-

FRANK VON HIPPEL, MARVIN MILLER, HAROLD FEIVESON, ANATOLI DIAKOV and FRANS BERKHOUT collaborate on issues of nuclear disarmament and nonproliferation. During the past five years, von Hippel, a physicist and professor of public and international affairs at Princeton University, has led an international research program on controlling both warheads and nuclear materials. Miller, a professor of nuclear engineering at the Massachusetts Institute of Technology, advises U.S. government agencies on nonproliferation policy. Feiveson is a senior research policy analyst at Princeton and editor of *Science & Global Security*. Diakov is director of the Center for Arms Control, Energy and Environmental Studies at the Moscow Institute of Physics and Technology. Berkhout, a research associate at Princeton's Center for Energy and Environmental Studies, analyzes issues related to the reprocessing of nuclear fuel from civilian reactors and the recycling of plutonium.

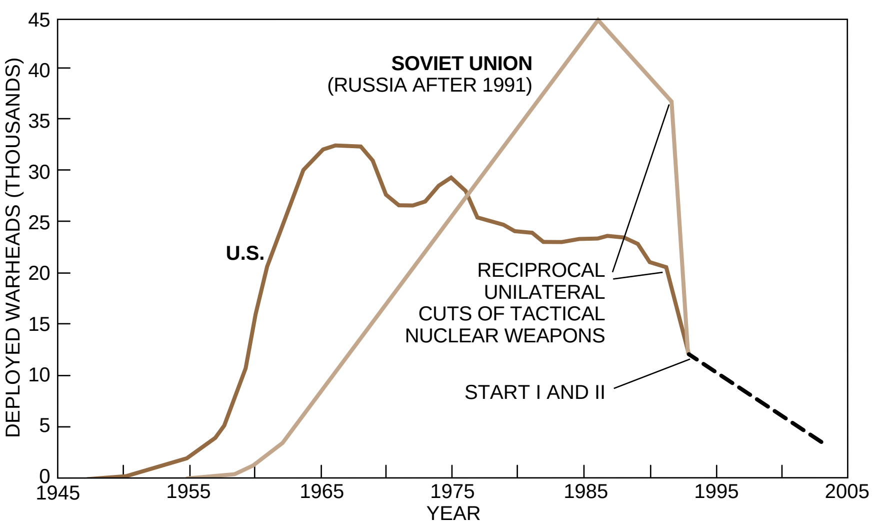
TACTICAL NUCLEAR WARHEAD from the former Soviet Union is loaded on board a truck in Ukraine for transport to Russia, where it is to be stored. The withdrawal of tactical warheads from service in 1992 eased nuclear tensions, but now the U.S. and the Soviet Union's successors must decide what to do with this warhead and tens of thousands more.

NUCLEAR WEAPONS of the former Soviet Union are scattered across the territory of four successor states. More than 3,000 remain in Ukraine, Kazakhstan and Belarus but should eventually be shipped to Russia for disposal. Warheads are currently being dismantled at four sites in Russia. Negotiations are under way to dilute at least 500 tons of the resulting highly enriched uranium with natural uranium and sell it to the U.S. for use as reactor fuel. Weapon-grade plutonium is still being separated from spent reactor fuel at facilities near Tomsk and Krasnoyarsk. A third plant, near Chelyabinsk, has separated more than 25 tons of civilian-grade plutonium from power-reactor fuel since 1978.

nounced in 1991 by President George Bush and Russian leader Mikhail S. Gorbachev, for decommissioning most Soviet and U.S. tactical warheads.

Although START I and START II will increase the scale of the warhead disposal problem, Russia and the U.S. are already dismantling nuclear warheads at a considerable rate—between 1,000 and 2,000 warheads a year in each country.

Taking a thermonuclear warhead apart safely is a technically demanding task. Most modern strategic warheads consist of a "primary" (fission) explosive and a thermonuclear (fusion) "secondary" that is ignited by the explosion of the primary. The hollow, spherical "pit" of the primary holds the warhead's plutonium, three to four kilograms on average, sometimes with some highly enriched uranium (that is, HEU, incorporating more than 90 percent chain-reacting uranium 235). The secondary generally also contains HEU, for a total of perhaps 15 kilograms for the average warhead. All told, surplus U.S. warheads contain about 50 tons of plutonium and up to 400 tons of HEU. Surplus Soviet warheads, including about 10,000 that have already been dismantled, contain about 100 tons of pluto-



SUPERPOWER ARSENALS have declined precipitously since 1991, when George Bush and Mikhail S. Gorbachev announced that most of their nations' tactical nuclear warheads would be placed in storage. Under current treaties, each nation is to reduce the number of strategic weapons deployed to between 3,000 and 3,500 by the year 2003. In the absence of further agreements, Russia and the U.S. will each retain a total of about 5,000 deployed strategic and tactical warheads. The agreements mandating these reductions, however, do not dictate what is to become of the warheads taken out of service or of the uranium and plutonium they contain.

nium and more than 500 tons of HEU.

When workers dismantle a warhead, they first remove the primary and secondary from the bomb casing and then detach the chemical explosives that surround the pit. Finally, they recover the plutonium and HEU for reuse or storage. In the U.S., disassembly takes place at the Department of Energy's Pantex facility near Amarillo, Tex. The secondaries go to the department's Y-12 plant in Oak Ridge, Tenn., where their uranium is recovered and stored.

Until 1989, U.S. pits went to the Energy Department's Rocky Flats plant near Denver, where their plutonium was recovered and purified for reuse. The plant was closed because of environmental and safety problems, however, and a replacement has yet to be built. In the meantime, pits are stored in sealed canisters in heavily protected bunkers at Pantex. There are 60 of these so-called igloos, each with room for up to about 400 pits, which is more than sufficient to accommodate the pits from all the U.S. warheads currently scheduled to be taken out of service.

In Russia, warheads are being dismantled at four sites with a reported combined disassembly capacity of up to 6,000 warheads a year. The Russian Ministry of Atomic Energy has asked for U.S. assistance to construct a secure central store for 40,000 containers for nuclear warhead components or materials near the Siberian city of Tomsk, one of Russia's three plutonium production centers. The Tomsk city government has opposed the plan because of concern about potential plutonium hazards. After the explosion that destroyed part of the nearby Tomsk-7 reprocessing plant this past April, the proposal was officially "deferred."

Whatever the fate of this facility, secure storage of nuclear materials is the most critical near-term objective for both Russia and the U.S. Such storage would protect materials until they can be processed into more proliferation-resistant forms. So long as the recovered nuclear materials remain in forms easily converted back to weapons, their existence will erode confidence in the disarmament process and raise dangers of diversion to nonnuclear nations or terrorists.

The obvious way to render highly enriched uranium useless for weapons is to blend it with large quantities of the non-chain-reacting uranium isotope, uranium 238, which makes up 99.3 percent of natural uranium. Reconstituting the enriched fraction requires isotope separation techniques, which have been mastered by only a few countries. If the HEU is diluted to about 4 percent uranium 235, the resulting "low-enriched" uranium can be used to fuel standard light-water nuclear power reactors.

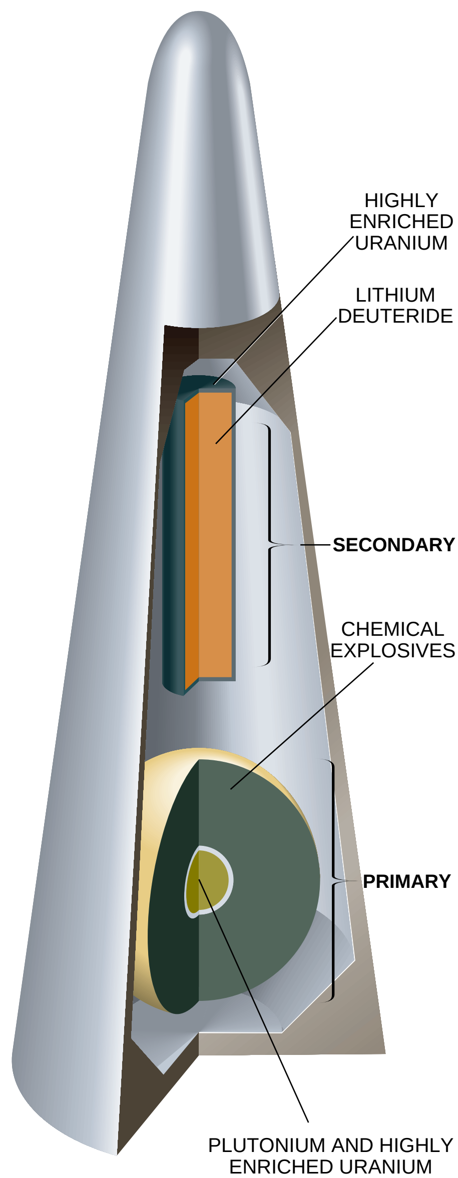
Indeed, following a suggestion by Thomas Neff of the Massachusetts Institute of Technology, the U.S. government has agreed to pay roughly $10 billion for low-enriched uranium derived from about 500 tons of weapon-grade uranium recovered from surplus Soviet warheads. This quantity could fuel about one eighth of the world's nuclear capacity during the 20-year period covered by the contract. According to present plans, the Russians will dilute the HEU in a facility near Ekaterinburg (formerly Sverdlovsk) before shipment to the U.S.

About 400 of the approximately 500 tons of weapons uranium in the U.S. stockpile will probably also become surplus. A few tons a year will be used to fuel nuclear-powered warships and submarines, as well as reactors devoted to research or to making radioisotopes for medical and other uses. The rest should be diluted down to low enrichment levels as quickly as possible and held for eventual sale as power-reactor fuel. This action would reduce the cost of safeguarding the material and would also reassure Russia and other countries that U.S. arms reductions are irreversible.

The 150 tons of surplus plutonium that dismantled warheads will yield poses a thornier problem because it cannot be denatured isotopically in the same way as weapons uranium. But reclaiming plutonium for reuse in weapons can be made much more difficult by mixing it with radioactive fission products. One obvious way to do this is to substitute the weapons plutonium for uranium 235 in so-called mixed-oxide fuel that can be used in commercial light-water reactors. Three years in a reactor core would reduce the amount of plutonium in the fuel by about 40 percent.

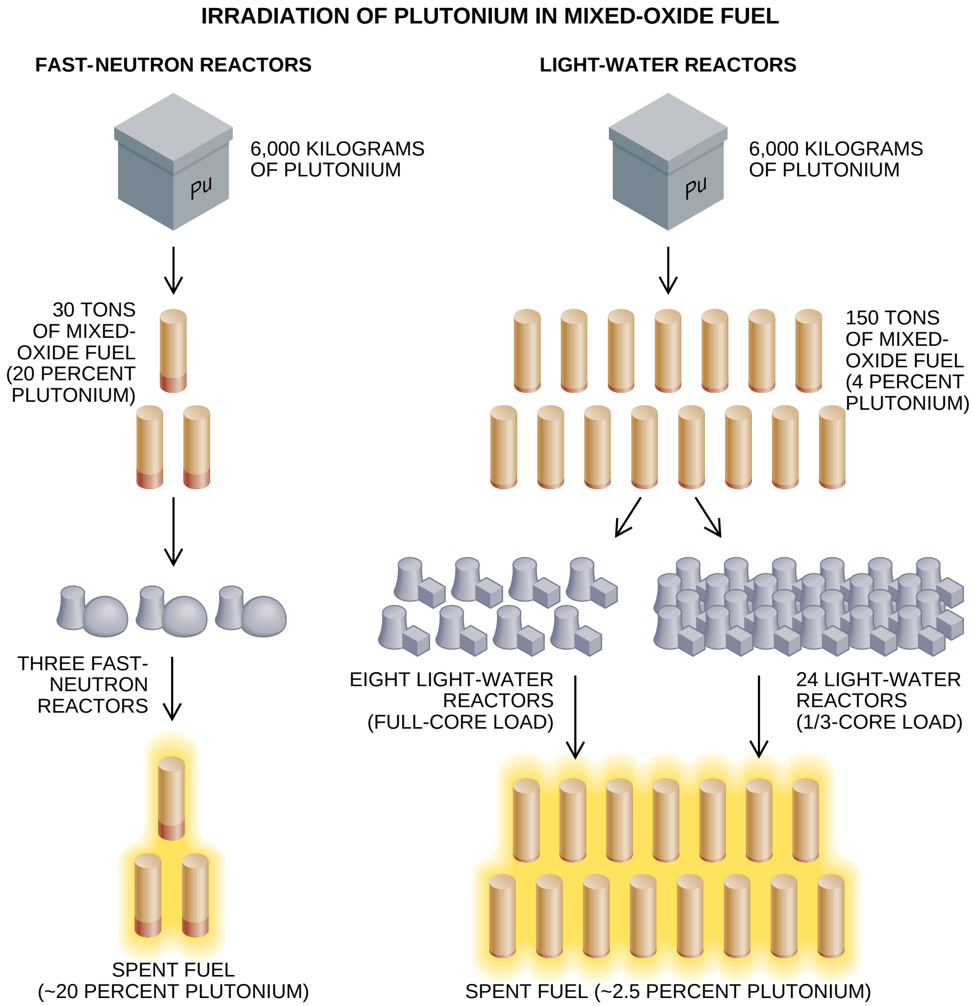
The plutonium remaining in the discharged spent fuel would have an increased fraction of plutonium isotopes other than plutonium 239 (the preferred isotope for warheads), making it less attractive as a weapons material. This reactor-grade plutonium, however, could still be separated and used to make simple bombs having yields of about 1,000 tons of high explosive. (To put this in perspective, the bomb that recently wreaked such havoc at the World Trade Center in New York City contained about half a ton of high explosive.)

Japan and some Western European nations have already set up a partial infrastructure for recycling plutonium recovered from spent power-reactor fuel, so the addition of weapons plutonium to this system might seem attractive. Unfortunately, the electric utilities in these countries have no interest in pursuing this option. The cost of manufacturing mixed-oxide fuel is currently considerably greater than the cost of low-enriched uranium fuel, and in any case, these nations already anticipate a significant surplus of civilian plutonium.



NUCLEAR WARHEAD typically consists of a fission "primary" and a fusion-fission "secondary." When weapons are dismantled, their chemical explosives are detached; the plutonium of the primary and the highly enriched uranium of the secondary are then removed for processing.

PLUTONIUM DISPOSAL is a problem that has yet to be definitively solved. Two solutions have been proposed. One would employ plutonium to fuel nuclear reactors, irradiating it and reducing its value for weapons. The other, safer and less costly, would incorporate the metal in glass "logs" soon to be manufactured for storing high-level radioactive waste.



Furthermore, mixed-oxide fuel raises serious security concerns because the freshly manufactured material contains plutonium in a readily separable form, unaccompanied by fission products. Such concerns led the U.S. to reject commercial plutonium recycling more than a decade ago. As a result, the U.S. has no facility for making mixed-oxide, light-water reactor fuel.

Russia also has no mixed-oxide fuel fabrication plant. Even if it did, the rate at which the plutonium could be irradiated in Russian light-water reactors would be very limited. Plutonium's nuclear characteristics limit the fraction of mixed-oxide fuel that can be substituted for low-enriched uranium in most light-water reactors to about one third of the core. Consequently, a 1,000-megawatt electric light-water reactor could process only about 300 kilograms of weapons plutonium a year. Russia has seven such reactors operating, with another nearly complete, and so could irradiate about 2.5 tons of plutonium a year. At this rate, it would take 40 years to irradiate Russia's 100 tons of surplus weapons plutonium. During this entire period, the plutonium in Russian fuel fabrication plants and in transit to power-reactor sites would be susceptible to diversion.

Security risks could be reduced by building reactors, designed to accept full cores of mixed-oxide fuel, at a single highly secured site in each country. Various reactor types have been proposed for this purpose. The one that could probably be built most quickly is a light-water reactor manufactured by ABB Combustion Engineering, which was specifically designed to be easily adaptable to a full plutonium core.

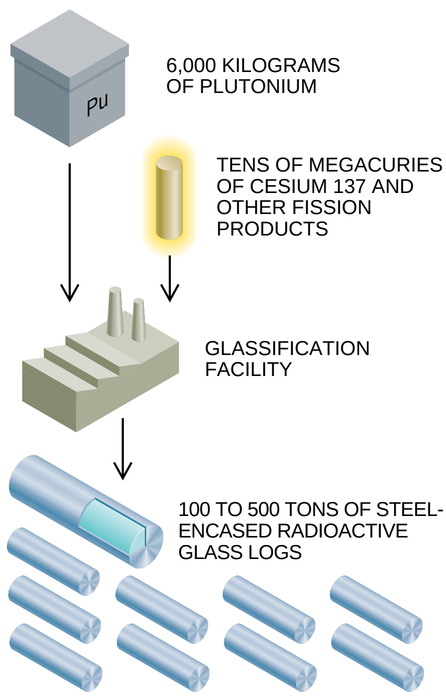
Other candidates include the liquid metal–cooled fast-neutron reactor and the high-temperature gas-cooled reactor; advanced versions of these concepts are under development in the U.S. and other countries. The fast-neutron reactor can irradiate more plutonium than can a light-water reactor of equivalent power because of the higher percentage of plutonium in the fuel. Unfortunately, without recycling, the plutonium in the spent fuel would still be near weapon grade. The gas-cooled reactor, in contrast, could irradiate plutonium to a point where most of it would be destroyed and the remainder rendered even more undesirable for weapons than the plutonium in spent fuel from a light-water reactor. Yet it makes little sense to pursue virtually complete fission of military plutonium in the absence of plans to treat similarly the much larger quantities of civilian plutonium (more than 1,000 tons by the turn of the century) now accumulating in unreprocessed power-reactor fuel.

Moreover, both the liquid metal–cooled and gas-cooled reactors require considerable development and demonstration before they can be considered ready for full-scale implementation. (This is even more true of another proposed route to plutonium elimination: irradiation by neutrons produced in targets bombarded by protons accelerated to high energies.) The cost would be several billion dollars and at least a decade of delay. And once the technology had been demonstrated, there would still be costly production facilities to build.

Given these difficulties, researchers in the U.S. and Russia are considering alternatives that could possibly be implemented more rapidly and cheaply. In particular, we and others have been examining the feasibility of disposing of plutonium together with radioactive waste. Facilities have already been constructed in both countries, as well as in France, Britain and Belgium, to dispose of high-level reprocessing waste by incorporating it into glass that will eventually be placed in deep geologic repositories. Although disposal of plutonium with radioactive waste would forgo the electricity it could generate, this loss is insignificant in the larger context. At present uranium and plutonium prices, plutonium will not be an economic fuel for at least several decades. In addition, one or two hundred tons of the metal could power the world's current nuclear capacity for only a fraction of a year.

The security threat posed by this material should therefore take precedence. Direct disposal of plutonium would involve much less handling and transport—and so less risk of diversion—than would its use in fuel. If the use of

MIXING OF PLUTONIUM WITH HIGH-LEVEL RADIOACTIVE WASTE



plutonium for reactor fuel proves economically and politically viable at some future time, there will still be thousands of tons of civilian plutonium recoverable from spent fuel.

A waste glassification plant has been built in Aiken, S.C., the site of the now defunct Savannah River military plutonium production complex. Between 1994 and 2009 this facility is expected to produce at least 8,000 tons of radioactive glass in the form of massive steel-sheathed "logs" three meters long and 0.6 meter in diameter, each containing about half a ton of high-level waste slurry mixed with 1.2 tons of borosilicate glass. Seventy tons of plutonium could be dissolved in these logs without raising the concentration to levels above those in spent power-reactor fuel.

It would take at least five years to complete the safety assessments and other preparations required for incorporating weapons plutonium into radioactive glass at Savannah River, but experts there have not identified any significant technical obstacles. Because the glass would be made in any case, the extra costs involved are those related to the preprocessing of the plutonium and its introduction into the melter and for appropriate safeguards and security arrangements. These costs would probably be less than those of irradiating plutonium in light-water reactors.

Although embedding the weapons plutonium in radioactive glass means that it would remain weapon grade, the highly radioactive fission products would make it at least as difficult to recover the plutonium from the glass logs as from spent fuel. The plutonium would be inaccessible to subnational groups, and even a determined country would need considerable time and resources to recover it.

Another possibility is to put the plutonium into logs without high-level waste, instead adding elements, such as gadolinium, that are very similar chemically to plutonium and thus difficult to separate from it. This strategy would make the plutonium inaccessible to subnational groups, even though a would-be nuclear nation could still recover it relatively easily. The plutonium-dilutant mixture could also be "spiked" with cesium 137, a fission product that is an intense gamma emitter and has a 30-year half-life.

Russia is glassifying high-level waste at its reprocessing site near Chelyabinsk. About as much waste resides in the Chelyabinsk tanks (measured in terms of its radioactivity) as at Savannah River, but the phosphate glass used at Chelyabinsk does not appear to be as durable as the borosilicate glass used in Western Europe, Japan and the U.S., nor does it have the safety advantages associated with the neutron-absorbing boron.

If borosilicate glassification technology were transferred to Russia, its weapons plutonium could easily be embedded in such glass. Unfortunately, the Russian nuclear establishment has shown little enthusiasm for glassification or, more generally, for processing plutonium into more diversion-resistant forms. This material was produced at enormous human and environmental cost; Russian nuclear officials consider it a national heritage. They prefer to store it for possible future use, even though safeguarding it for decades will be expensive and risky. A recognition of these costs and risks by the Russian political authorities, together with financial incentives and the knowledge that the U.S. is willing to render its own weapons plutonium inaccessible, may convince Russia to abandon its deadly treasure.

Assuming that the U.S. and former Soviet states can come to an agreement on how to dispose of surplus warheads, there is still the question of verification. International confidence in the nuclear-arms reduction process would be enhanced if disposal of surplus warheads could be subjected to outside monitoring. Moreover, experts at Los Alamos National Laboratory and at Pantex have concluded that effective monitoring could be carried out without revealing sensitive nuclear-warhead design information. Nevertheless, the U.S. government continues to pursue an essentially unilateral policy by limiting itself to the monitoring rights it can negotiate in connection with purchases of Soviet highly enriched uranium and assistance in building storage facilities for surplus weapons.

In addition, we believe the U.S. and Russia should conduct such monitoring on a bilateral basis through the warhead dismantlement stage, putting recovered uranium and plutonium under international safeguards after they have been processed to remove weapons design information. The International Atomic Energy Agency has already offered to monitor the storage and subsequent use or disposal of the surplus warhead materials. This combination of bilateral and international safeguards would help ensure that the dismantlement process was secure and that the nuclear materials would never be reused in weapons.

Russia's current leadership has indicated that it is agreeable to such comprehensive monitoring—if it is done on a reciprocal basis. It is not clear how long this window of opportunity will stay open. The U.S. should move quickly to offer Russia a reciprocal right to monitor U.S. warhead elimination. Ultimately, these steps should be reinforced by a strengthened nonproliferation regime in which production of weapons-usable materials is ended worldwide, not just in the U.S. and the former Soviet Union. Such a production ban would assure that reductions in existing nuclear arsenals are irreversible and would minimize the risk that other nations or terrorist groups will acquire the wherewithal to make nuclear weapons.

FURTHER READING

REVERSING THE ARMS RACE: HOW TO ACHIEVE AND VERIFY DEEP REDUCTIONS IN THE NUCLEAR ARSENALS. Edited by Frank von Hippel and Roald Z. Sagdeev. Gordon and Breach Science Publishers, 1990.

DISPOSITION OF SEPARATED PLUTONIUM. Frans Berkhout, Anatoli Diakov, Harold Feiveson, Helen Hunt, Edwin Lyman, Marvin Miller and Frank von Hippel in *Science & Global Security,* Vol. 3, Nos. 3–4, pages 161–214; March 1993.

ON THE APPLICATION OF IAEA SAFEGUARDS TO PLUTONIUM AND HIGHLY ENRICHED URANIUM FROM MILITARY INVENTORIES. Thomas E. Shea in *Science & Global Security,* Vol. 3, Nos. 3–4, pages 223–236; March 1993.

RUSSIAN/SOVIET NUCLEAR WARHEAD PRODUCTION. Thomas B. Cochran and Robert S. Norris. Natural Resources Defense Council, 1993.

Faster than Light?

Experiments in quantum optics show that two distant events can influence each other faster than any signal could have traveled between them

by Raymond Y. Chiao, Paul G. Kwiat and Aephraim M. Steinberg

For experimentalists studying quantum mechanics, the fantastic often turns into reality. A recent example emerges from the study of a phenomenon known as nonlocality, or "action at a distance." This concept calls into question one of the most fundamental tenets of modern physics, the proposition that nothing travels faster than the speed of light.

An apparent violation of this proposition occurs when a particle at a wall vanishes, only to reappear—almost instantaneously—on the other side. A reference to Lewis Carroll may help here. When Alice stepped through the looking glass, her movement constituted in some sense action at a distance, or nonlocality: her effortless passage through a solid object was instantaneous. The particle's behavior is equally odd. If we attempted to calculate the particle's average velocity, we would find that it exceeded the speed of light.

Is this possible? Can one of the most famous laws of modern physics be breached with impunity? Or is there something wrong with our conception of quantum mechanics or with the idea of a "traversal velocity"? To answer such questions, we and several other workers have recently conducted many optical experiments to investigate some of the manifestations of quantum nonlocality. In particular, we focus on three demonstrations of nonlocal effects. In the first example, we "race" two photons, one of which must move through a "wall." In the second instance, we look at how the race is timed, showing that each photon travels along the two different race paths simultaneously. The final experiment reveals how the simultaneous behavior of photon twins is coupled, even if the twins are so far apart that no signal has time to travel between them.

The distinction between locality and nonlocality is related to the concept of a trajectory. For example, in the classical world a rolling croquet ball has a definite position at every moment. If each moment is captured as a snapshot and the pictures are joined, they form a smooth, unbroken line, or trajectory, from the player's mallet to the hoop. At each point on this trajectory, the croquet ball has a definite speed, which is related to its kinetic energy. If it travels on a flat pitch, it rolls to its target. But if the ball begins to roll up a hill, its kinetic energy is converted into potential energy. As a result, it slows—eventually to stop and roll back down. In the jargon of physics such a hill is called a barrier, because the ball does not have enough energy to travel over it, and, classically, it always rolls back. Similarly, if Alice were unable to hit croquet balls (or rolled-up hedgehogs, as Carroll would have them) with enough energy to send them crashing through a brick wall, they would merely bounce off.

According to quantum mechanics, this concept of a trajectory is flawed. The position of a quantum mechanical particle, unlike that of a croquet ball, is not described as a precise mathematical point. Rather the particle is best represented as a smeared-out wave packet. This packet can be seen as resembling the shell of a tortoise, because it rises from its leading edge to a certain height and then slopes down again to its trailing edge. The height of the wave at a given position along this span indicates the probability that the particle occupies that position: the higher a given part of the wave packet, the more likely the particle is located there. The width of the packet from front to back represents the intrinsic uncertainty of the particle's location [*see box on page 57*]. When the particle is detected at one point, however, the entire wave packet disappears. Quantum mechanics does not tell us where the particle has been before this moment.

This uncertainty in location leads to one of the most remarkable consequences of quantum mechanics. If the hedgehogs are quantum mechanical, then the uncertainty of position permits the beasts to have a very small but perfectly real chance of appearing on the far side of the wall. This process is known as tunneling and plays a major role in science and technology. Tunneling is of central importance in nuclear fusion, certain high-speed electronic devices, the highest-resolution microscopes in existence and some theories of cosmology.

In spite of the name "tunneling," the barrier is intact at all times. In fact, if a particle were inside the barrier, its kinetic energy would be negative. Velocity is proportional to the square root of the kinetic energy, and so in the tunneling case one must take the square

RAYMOND Y. CHIAO, PAUL G. KWIAT and AEPHRAIM M. STEINBERG have been using nonlinear optics to study several fundamental features of quantum mechanics—namely, interference, nonlocality and tunneling. As an undergraduate at Princeton University, Chiao was directed by John A. Wheeler to quantize gravity. Despite his failure in this monumental task, Chiao received his bachelor's degree in 1961. He received his Ph.D. from the Massachusetts Institute of Technology under the tutelage of Charles Townes and since 1967 has been a professor of physics at the University of California, Berkeley. A fellow of the American Physical Society, Chiao is described by his students as a concert-quality pianist to within experimental error. Kwiat is a postdoctoral fellow studying quantum optics at Innsbruck University. He received his B.S. from M.I.T. in 1987 and recently earned his Ph.D. under Chiao's direction. He is also devoted to the study of aikido, a Japanese martial art. Steinberg received his B.S. in physics from Yale University in 1988. He worked at the École Normale Supérieure for one year before becoming a Ph.D. student of Chiao's. He spends most of each day doing physics and wishing he had more time to ski.

root of a negative number. Hence, it is impossible to ascribe a real velocity to the particle in the barrier. This is why when looking at the watch it has borrowed from the White Rabbit, the hedgehog that has tunneled to the far side of the wall wears—like most physicists since the 1930s—a puzzled expression. What time does the hedgehog see? In other words, how long did it take to tunnel through the barrier?

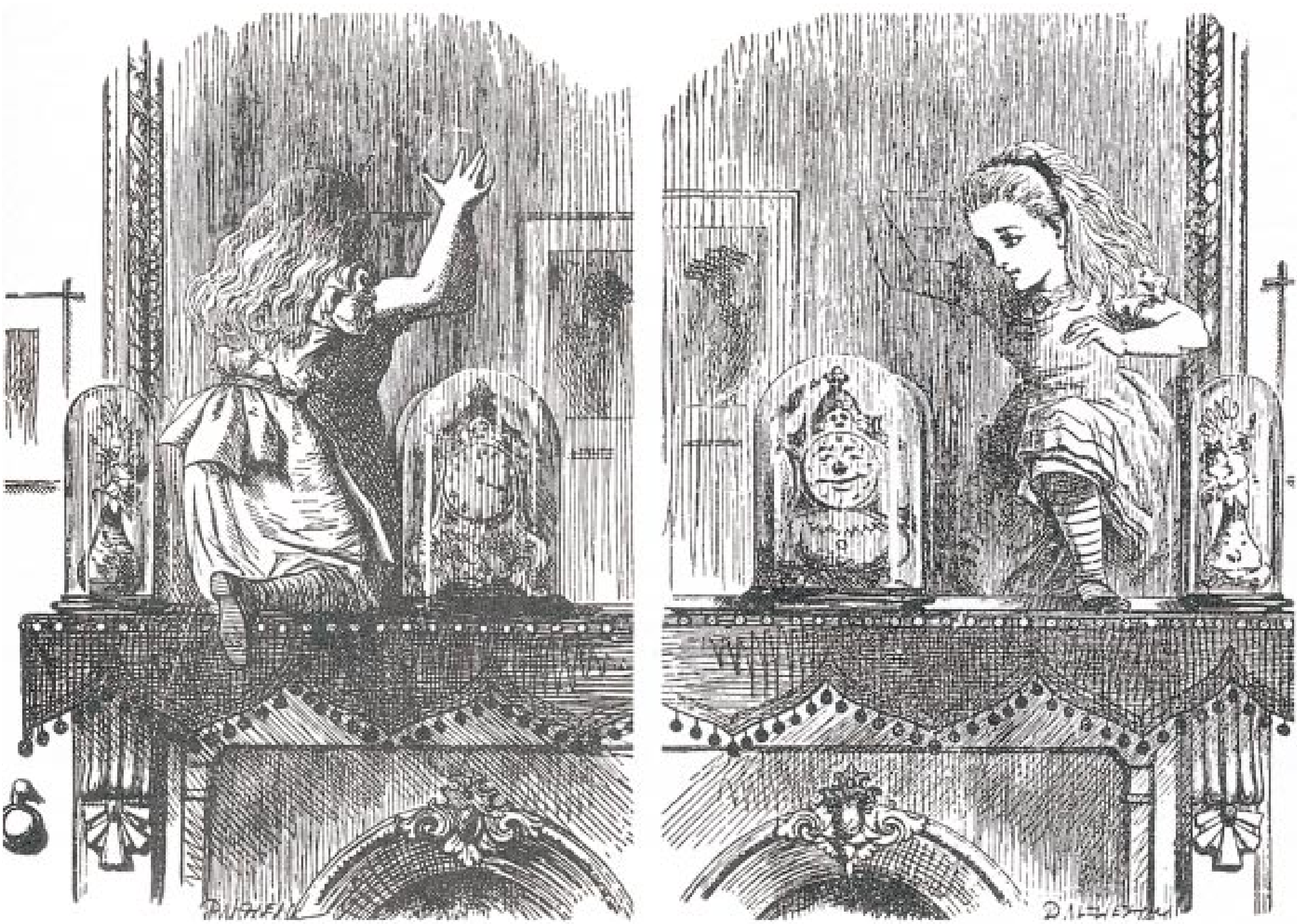
Over the years, many attempts have been made to answer the question of the tunneling time, but none has been universally accepted. Using photons rather than hedgehogs, our group has recently completed an experiment that provides one concrete definition of this time.

Photons are the elementary particles from which all light is made; a typical light bulb emits more than 100 billion such particles in one billionth of a second. Our experiment does not need nearly so many of them. To make our measurements, we used a light source that emits a pair of photons simultaneously. Each photon travels toward a different detector. A barrier is placed in the path of one of these photons, whereas the other is allowed to fly unimpeded. Most of the time, the first photon bounces off the barrier and is lost; only its twin is detected. Occasionally, however, the first photon tunnels through the barrier, and both photons reach their respective detectors. In this situation, we can compare their arrival times and thus see how long the tunneling process took.

The role of the barrier was played by a common optical element: a mirror. This mirror, however, is unlike the ordinary household variety (which relies on metallic coating and absorbs as much as 15 percent of the incident light). The laboratory mirrors consist of thin, alternating layers of two different types of transparent glass, through which light travels at slightly different speeds. These layers act as periodic "speed bumps." Individually, they would do little more than slow the light down. But when taken together and spaced appropriately, they form a region in which light finds it essentially impossible to travel. A multilayer coating one micron thick—one one-hundredth of the diameter of a typical human hair—reflects 99 percent of incident light at the photon energy (or, equivalently, the color of the light) for which it is designed. Our experiment looks at the remaining 1 percent of the photons, which tunnel through this looking glass.

During several days of data collection, more than one million photons tunneled through the barrier, one by one. We compared the arrival times for tunneling photons and for photons that had been traveling unimpeded at the speed of light. (The speed of light is so great that conventional electronics are hundreds of thousands of times too slow to perform the timing; the technique we used will be described later, as a second example of quantum nonlocality.)

The surprising result: on average, the tunneling photons arrived before those that traveled through air, implying an average tunneling velocity of about 1.7 times that of light. The result appears to contradict the classical notion of causality, because, according to Einstein's theory of relativity, no signal can travel faster than the speed of light. If signals



"TUNNELING" ALICE moves effortlessly through a mirror, much as photons do in experiments in quantum optics. Although he was not a physicist, Lewis Carroll almost seems to have anticipated a thorny 20th-century physics problem—that of the tunneling time—when he had Sir John Tenniel draw a strange face on the looking-glass clock.

LOOKING-GLASS CROQUET has Alice hitting rolled-up hedgehogs, each bearing an uncanny resemblance to a young Werner Heisenberg, toward a wall. Classically, the hedgehogs always bounce off. Quantum mechanically, however, a small probability exists that a hedgehog will appear on the far side. The puzzle facing quantum physicists: How long does it take

could move faster, effects could precede causes from the viewpoints of certain observers. For example, a light bulb might begin to glow before the switch was thrown.

The situation can be stated more precisely. If at some definite time you made a decision to start firing photons at a mirror by opening a starting gate, and someone else sat on the other side of the mirror looking for photons, how much time would elapse before the other person knew you had opened the gate? At first, it might seem that since the photon tunnels faster than light she would see the light before a signal traveling at the theoretical speed limit could have reached her, in violation of the Einsteinian view of causality. Such a state of affairs seems to suggest an array of extraordinary, even bizarre communication technologies. Indeed, the implications of faster-than-light influences led some physicists in the early part of the century to propose alternatives to the standard interpretation of quantum mechanics.

Is there a quantum mechanical way out of this paradox? Yes, there is, although it deprives us of the exciting possibility of toying with cause and effect. Until now, we have been talking about the tunneling velocity of photons in a classical context, as if it were a directly observable quantity. The Heisenberg uncertainty principle, however, indicates that it is not. The time of emission of a photon is not precisely defined, so neither is its exact location or velocity. In truth, the position of a photon is more correctly described by a bell-shaped probability distribution—our tortoise shell—whose width corresponds to the uncertainty of its location.

A relapse into metaphor might help to explain the point. The nose of each tortoise leaves the starting gate the instant of opening. The emergence of the nose marks the earliest time at which there is any possibility for observing a photon. No signal can ever be received before the nose arrives. But because of the uncertainty of the photon's location, on average a short delay exists before the photon crosses the gate. Most of the tortoise (where the photon is more likely to be detected) trails behind its nose.

For simplicity, we label the probability distribution of the photon that travels unimpeded to the detector as "tortoise 1" and that of the photon that tunnels as "tortoise 2." When tortoise 2 reaches the tunnel barrier, it splits into two smaller tortoises: one that is reflected back toward the start and one that crosses the barrier. These two partial tortoises together represent the probability distribution of a single photon. When the photon is detected at one position, its other partial tortoise instantly disappears. The reflected tortoise is bigger than the tunneling tortoise simply because the chances of reflection are greater than that of transmission (recall that the mirror reflects a photon 99 percent of the time).

We observe that the peak of tortoise 2's shell, representing the most likely position of the tunneling photon, reaches the finish line before the peak of tortoise 1's shell. But tortoise 2's nose arrives no earlier than the nose of tortoise 1. Because the tortoises' noses travel at the speed of light, the photon that signals the opening of the starting gate can never arrive earlier than the time allowed by causality [*see illustration on opposite page*].

In a typical experiment, however, the nose represents a region of such low probability that a photon is rarely observed there. The whereabouts of the photon, detected only once, are best predicted by the location of the peak. So even though the tortoises are nose and nose at the finish, the peak of tor-

to go through the wall? Does the traversal time violate Albert Einstein's famous speed limit?

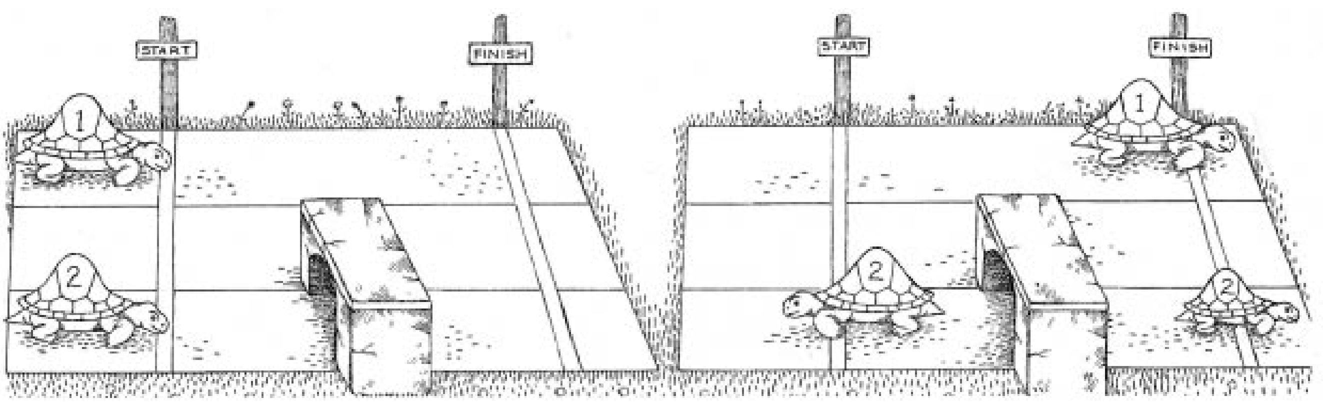
toise 2's shell precedes that of tortoise 1's (remember, the transmitted tortoise is smaller than tortoise 1). A photon tunneling through the barrier is therefore most likely to arrive before a photon traveling unimpeded at the speed of light. Our experiment confirmed this prediction.

But we do not believe that any individual part of the wave packet moves faster than light. Rather the wave packet gets "reshaped" as it travels, until the peak that emerges consists primarily of what was originally in front. At no point does the tunneling-photon wave packet travel faster than the free-traveling photon. In 1982 Steven Chu of Stanford University and Stephen Wong, then at AT&T Bell Laboratories, observed a similar reshaping effect. They experimented with laser pulses consisting of many photons and found that the few photons that made it through an obstacle arrived sooner than those that could move freely. One might suppose that only the first few photons of each pulse were "allowed" through and thus dismiss the reshaping effect. But this interpretation is not possible in our case, because we study one photon at a time. At the moment of detection, the entire photon "jumps" instantly into the transmitted portion of the wave packet, beating its twin to the finish more than half the time.

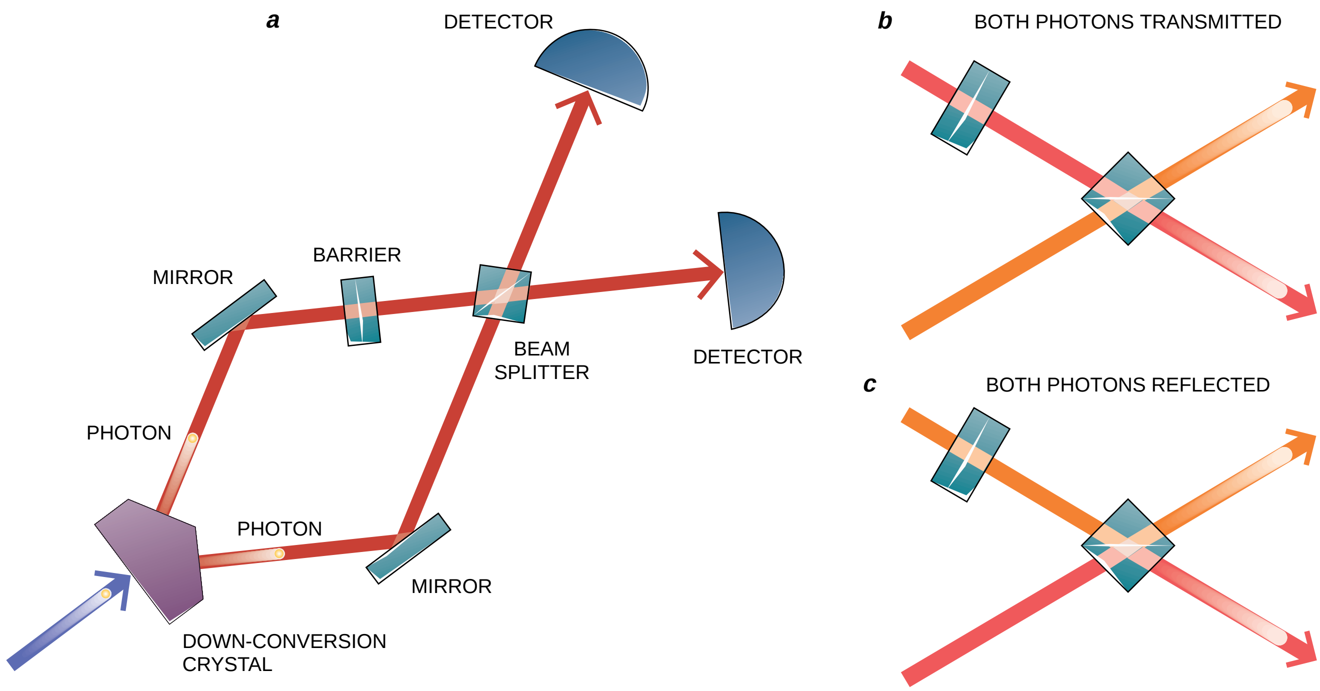
Although reshaping seems to account for our observations, the question still lingers as to why reshaping should occur in the first place. No one yet has any physical explanation for the rapid tunneling. In fact, the question had puzzled investigators as early as the 1930s, when physicists such as Eugene Wigner of Princeton University had noticed that quantum theory seemed to imply such high tunneling speeds. Some assumed that approximations used in that prediction must be incorrect, whereas others held that the theory was correct but required cautious interpretation. Some researchers, in particular Markus Büttiker and Rolf Landauer of the IBM Thomas J. Watson Research Center, suggest that quantities other than the arrival time of the wave packet's peak (for example, the angle through which a "spinning" particle rotates while tunneling) might be more appropriate for describing the time "spent" inside the barrier. Although quantum mechanics can predict a particle's average arrival time, it lacks the classical notion of trajectories, without which the meaning of time spent in a region is unclear.

One hint to explain fast tunneling time stems from a peculiar characteristic of the phenomenon. According to theory, an increase in the width of the barrier does not lengthen the time needed by the wave packet to tunnel through. This observation can be roughly understood using the uncertainty principle. Specifically, the less time we spend studying a photon, the less certain we can be of its energy. Even if a photon fired at a barrier does not have enough energy to cross it, in some sense a short period initially exists during which the particle's energy is uncertain. During this time, it is as though the photon could temporarily borrow enough extra energy to make it across the barrier. The length of this grace period depends only on the borrowed energy, not on the width of the barrier. No matter how wide the barrier becomes, the transit time across it remains the same. For a sufficiently wide barrier, the apparent traversal speed would exceed the speed of light.

Obviously, for our measurements to be meaningful, our tortoises had to run exactly the same distance. In essence, we had to straighten the racetrack so that neither tortoise had the advantage of the inside lane. Then, when we placed a barrier in one path, any delay or acceleration would be attributed solely to quantum tunneling. One way to set up two equal lanes would be to determine how much time



RACING TORTOISES help to characterize tunneling time. Each represents the probability distribution of the position of a photon. The peak is where a photon is most likely to be detected. The tortoises start together (*left*). Tortoise 2 encounters a barrier and splits in two (*right*). Because the chance of tunneling is low, the transmitted tortoise is small, whereas the reflected one is nearly as tall as the original. On those rare occasions of tunneling, the peak of tortoise 2's shell on average crosses the finish line first—implying an average tunneling velocity of 1.7 times the speed of light. But the tunneling tortoise's nose never travels faster then light—note that both tortoises remain "nose and nose" at the end. Hence, Einstein's law is not violated.



TWIN-PHOTON INTERFEROMETER (*a*) precisely times racing photons. The photons are born in a down-conversion crystal and are directed by mirrors to a beam splitter. If one photon beats the other to the beam splitter (because of the barrier), both detectors will be triggered in about half the races. Two possibilities lead to such coincidence detections: both photons are transmitted by the beam splitter (*b*), or both are reflected (*c*). Aside from their arrival times, there is no way of determining which photon took which route; either could have traversed the barrier. (This nonlocality actually sustains the performance of the interferometer.) If both photons reach the beam splitter simultaneously, for quantum reasons they will head in the same direction, so that both detectors do not go off. The two possibilities shown are then said to interfere destructively.

it takes for a photon to travel from the source to the detector for each path. Once the times were equal, we would know the paths were also equal.

But performing such a measurement with a conventional stopwatch would require one whose hands went around nearly a billion billion times per minute. Fortunately, Leonard Mandel and his co-workers at the University of Rochester have developed an interference technique that can time our photons.

Mandel's quantum stopwatch relies on an optical element called a beam splitter [*see illustration above*]. Such a device transmits half the photons striking it and reflects the other half. The racetrack is set up so that two photon wave packets are released at the same time from the starting gate and approach the beam splitter from opposite sides. For each pair of photons, there are four possibilities: both photons might pass through the beam splitter; both might rebound from the beam splitter; both could go off together to one side; or both could go off together to the other side. The first two possibilities—that both photons are transmitted or both reflected—result in what are termed coincidence detections. Each photon reaches a different detector (placed on either side of the beam splitter), and both detectors are triggered within a billionth of a second of each other. Unfortunately, this time resolution is about how long the photons take to run the entire race and hence is much too coarse to be useful.

So how do the beam splitter and the detectors help in the setup of the racetrack? We simply tinker with the length of one of the paths until all coincidence detections disappear. By doing so, we make the photons reach the beam splitter at the same time, effectively rendering the two racing lanes equal. Admittedly, the proposition sounds peculiar—after all, equal path lengths would seem to imply coincident arrivals at the two detectors. Why would the absence of such events be the desired signal?

The reason lies in the way quantum mechanical particles interact with one another. All particles in nature are either bosons or fermions. Identical fermions (electrons, for example) obey the Pauli exclusion principle, which prevents any two of them from ever being in the same place at the same time. In contrast, bosons (such as photons) like being together. Thus, after reaching the beam splitter at the same time, the two photons prefer to head in the same direction. This preference leads to the detection of fewer coincidences (none, in an ideal experiment) than would be the case if the photons acted independently or arrived at the beam splitter at different times.

Therefore, to make sure the photons are in a fair race, we adjust one of the path lengths. As we do this, the rate of coincident detections goes through a dip whose minimum occurs when the photons take exactly the same amount of time to reach the beam splitter. The width of the dip (which is the limiting factor in the resolution of our experiments) corresponds to the size of the photon wave packets—typically, about the distance light moves in a few hundredths of a trillionth of a second.

Only when we knew that the two path lengths were equal did we install the barrier and begin the race. We then found that the coincidence rates were no longer at a minimum, implying that one of the photons was reaching the beam splitter first. To restore the minimum, we had to lengthen the path taken by the tunneling photon. This correction indicates that photons take less time to cross a barrier than to travel in air.

Even though investigators designed racetracks for photons and a clever timekeeping device for the race, the competition still should have been difficult to conduct. The fact that the test could be carried out at all constitutes a second validation of the principle of nonlocality, if not for which precise timing of the race would have been impossible. To determine the emission time of a photon most precisely, one would obviously like the photon wave packets to be as short as possible. The uncertainty principle, however, states that the more accurately one determines the emission time of a photon, the more uncertainty one has to accept in knowing its energy, or color [*see box below*].

Because of the uncertainty principle, a fundamental trade-off should emerge in our experiments. The colors that make up a photon will disperse in any kind of glass, widening the wave packet and reducing the precision of the timing. Dispersion arises from the fact that different colors travel at various speeds in glass—blue light generally moves more slowly than red. A familiar example of dispersion is the splitting of white light into its constituent colors by a prism.

As a short pulse of light travels through a dispersive medium (the barrier itself or one of the glass elements used to steer the light), it spreads out into a "chirped" pulse: the redder part pulls ahead, and the bluer hues lag behind [*see illustration on next page*]. A simple calculation shows that the width of our photon pulses would quadruple on passage through an inch of glass. The presence of such broadening should have made it well nigh impossible to tell which tortoise crossed the finish line first. Remarkably, the widening of the photon pulse did not degrade the precision of our timing.

Herein lies our second example of quantum nonlocality. Essentially both twin photons must be traveling both paths simultaneously. Almost magically, potential timing errors cancel out as a result.

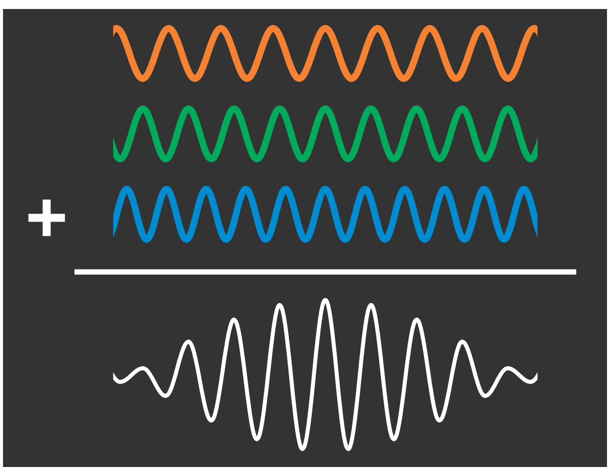
To understand this cancellation effect, we need to examine a special property of our photon pairs. The pairs are born in what physicists call "spontaneous parametric down-conversion." The process occurs when a photon travels into a crystal that has nonlinear optical properties. Such a crystal can absorb a single photon and emit a pair of other photons, each with about half the energy of the parent, in its place (this is the meaning of the phrase "down-conversion"). An ultraviolet photon, for instance, would produce two infrared ones. The two photons are emitted simultaneously, and the sum of their energies exactly equals the energy of the parent photon. In other words, the colors of the photon pairs are correlated—if one is slightly bluer (and thus travels more slowly in glass), then the other must be slightly redder (and must travel more quickly).

One might think that differences between siblings might affect the outcome of the race—one tortoise might be more athletic than the other. Yet because of nonlocality, any discrepancy between the pair proves irrelevant. The key point is that neither detector has any way of identifying which of the photons took which path. Either photon

Wave Packets

A good way to understand wave packets is to construct one, by adding together waves of different frequencies. We start with a central frequency (*denoted by the green curve*), a wave with no beginning and no end. If we now add two more waves of slightly lower and higher frequency (*orange and blue curves, respectively*), we obtain a pulselike object (*white curve*). When enough frequencies are added, a true pulse, or wave packet, can be formed, which is confined to a small region of space. If the range of frequencies used to make the pulse were decreased (for example, by using colors only from yellow to green, instead of from orange to blue), we would create a longer pulse. Conversely, if we had included all colors from red to violet, the packet could have been even shorter.

Mathematically speaking, if we use $\Delta\nu$ for the width of the range of colors and Δt for the duration of the pulse, then we can write



$$\Delta\nu \, \Delta t \geq 1/4\pi,$$

which simply expresses the fact that a wider color range is needed to make a shorter wave packet. It holds true for any kind of wave—light, sound, water and so on.

The phenomenon acquires new physical significance when one makes the identification of electromagnetic frequency, ν, with photon energy, E, via the Planck-Einstein relation $E = h\nu$, where h is Planck's constant. The particle aspect of quantum mechanics enters at this point. In other words, a photon's energy depends on its color. Red photons have about three fifths the energy of blue ones. The above mathematical expression can then be rewritten as

$$\Delta E \, \Delta t \geq h/4\pi.$$

Physicists have become so attached to this formula that they have named it: Heisenberg's uncertainty principle. (An analogous and perhaps more familiar version exists for position and momentum.) One consequence of this principle for the experiments described in the article is that it is strictly impossible, even with a perfect apparatus, to know precisely both the time of emission of a photon and its energy.

Although we arrived at the uncertainty principle by considering the construction of wave packets, its application is remarkably far more wide-reaching and its connotations far more general. We cannot overemphasize that the uncertainty is inherent in the laws of nature. It is not merely a result of inaccurate measuring devices in our laboratories. The uncertainty principle is what keeps electrons from falling into the atomic nucleus, ultimately limits the resolution of microscopes and, according to some astrophysical theories, was initially responsible for the nonuniform distribution of matter in the universe.

might have passed through the barrier.

Having two or more coexisting possibilities that lead to the same final outcome results in what is termed an interference effect. Here each photon takes both paths simultaneously, and these two possibilities interfere with each other. That is, the possibility that the photon that went through the glass was the redder (faster) one interferes with the possibility that it was the bluer (slower) one. As a result, the speed differences balance, and the effects of dispersion cancel out. The dispersive widening of the individual photon pulses is no longer a factor. If nature acted locally, we would have been hard-pressed to conduct any measurements. The only way to describe what happens is to say that each twin travels through both the path with the barrier and the free path, a situation that exemplifies nonlocality.

Thus far we have discussed two nonlocal results from our quantum experiments. The first is the measurement of tunneling time, which requires two photons to start a race at exactly the same time. The second is the dispersion cancellation effect, which relies on a precise correlation of the racing photons' energies. In other words, the photons are said to be correlated in energy (what they do) and in time (when they do it). Our final example of nonlocality is effectively a combination of the first two. Specifically, one photon "reacts" to what its twin does instantaneously, no matter how far apart they are.

Knowledgeable readers may protest at this point, claiming that the Heisenberg uncertainty principle forbids precise specification of both time and energy. And they would be right, for a single particle. For two particles, however, quantum mechanics allows us to define simultaneously the difference between their emission times and the sum of their energies, even though neither particle's time or energy is specified. This fact led Einstein, Boris Podolsky and Nathan Rosen to conclude that quantum mechanics is an incomplete theory. In 1935 they formulated a thought experiment to demonstrate what they believed to be the shortcomings of quantum mechanics.

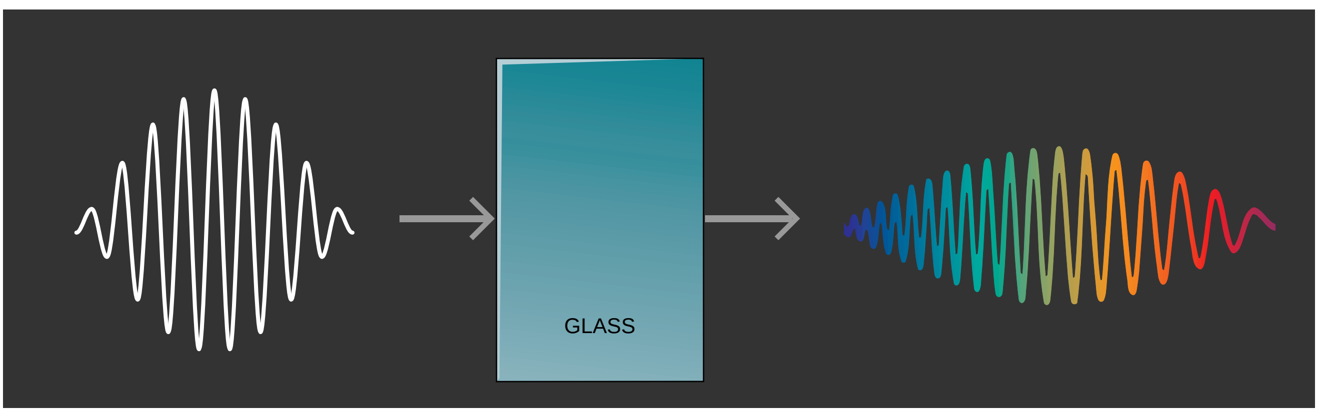
If one believes quantum mechanics, the dissenting physicists pointed out, then any two particles produced by a process such as down-conversion are coupled. For example, suppose we measure the time of emission of one particle. Because of the tight time correlation between them, we could predict with certainty the emission time of the other particle, without ever disturbing it. We could also measure directly the energy of the second particle and then infer the energy of the first particle. Somehow we would have managed to determine precisely both the energy and the time of each particle—in effect, beating the uncertainty principle. How can we understand the correlations and resolve this paradox?

There are basically two options. The first is that there exists what Einstein called "spooklike actions at a distance" (*spukhafte Fernwirkungen*). In this scenario, the quantum mechanical description of particles is the whole story. No particular time or energy is associated with any photon until, for example, an energy measurement is made. At that point, only one energy is observed. Because the energies of the two photons sum to the definite energy of the parent photon, the previously undetermined energy of the twin photon, which we did not measure, must instantaneously jump to the value demanded by energy conservation. This nonlocal "collapse" would occur no matter how far away the second photon had traveled. The uncertainty principle is not violated, because we can specify only one variable or the other: the energy measurement disrupts the system, instantaneously introducing a new uncertainty in the time.

Of course, such a crazy, nonlocal model should not be accepted if a simpler way exists to understand the correlations. A more intuitive explanation is that the twin photons leave the source at definite, correlated times, carrying definite, correlated energies. The fact that quantum mechanics cannot specify these properties simultaneously would merely indicate that the theory is incomplete.

Einstein, Podolsky and Rosen advocated the latter explanation. To them, there was nothing at all nonlocal in the observed correlations between particle pairs, because the properties of each particle are determined at the moment of emission. Quantum mechanics was only correct as a probabilistic theory, a kind of photon sociology, and could not completely describe all individual particles. One might imagine that there exists an underlying theory that could predict the specific results of all possible measurements and show that particles act locally. Such a theory would be based on some hidden variable yet to be discovered. In 1964 John S. Bell of CERN, the European laboratory for particle physics near Geneva, established a theorem showing that all invocations of local, hidden variables give predictions different from those stated by quantum mechanics.

Since then, experimental results have supported the nonlocal (quantum mechanical) picture and contradicted the intuitive one of Einstein, Podolsky and Rosen. Much of the credit for the pioneering work belongs to the groups led by John Clauser of the University of California at Berkeley and Alain Aspect, now at the Institute of Optics in Orsay. In the 1970s and



DISPERSION of a light pulse occurs because each color travels at a different speed. A short light pulse passing through a piece of glass will broaden into a "chirped" wave packet: the redder colors pull ahead while the bluer hues lag behind.

early 1980s they examined the correlations between polarizations in photons. The more recent work of John G. Rarity and Paul R. Tapster of the Royal Signals and Radar Establishment in England explored correlations between the momentum of twin photons. Our group has taken the tests one step further. Following an idea proposed by James D. Franson of Johns Hopkins University in 1989, we have performed an experiment to determine whether some local hidden variable model, rather than quantum mechanics, can account for the energy and time correlations.

In our experiment, photon twins from our down-conversion crystal are separately sent to identical interferometers [*see illustration at right*]. Each interferometer is designed much like an interstate highway with an optional detour. A photon can take a short path, going directly from its source to its destination. Or it can take the longer, detour path (whose length we can adjust) by detouring through the rest station before continuing on its way.

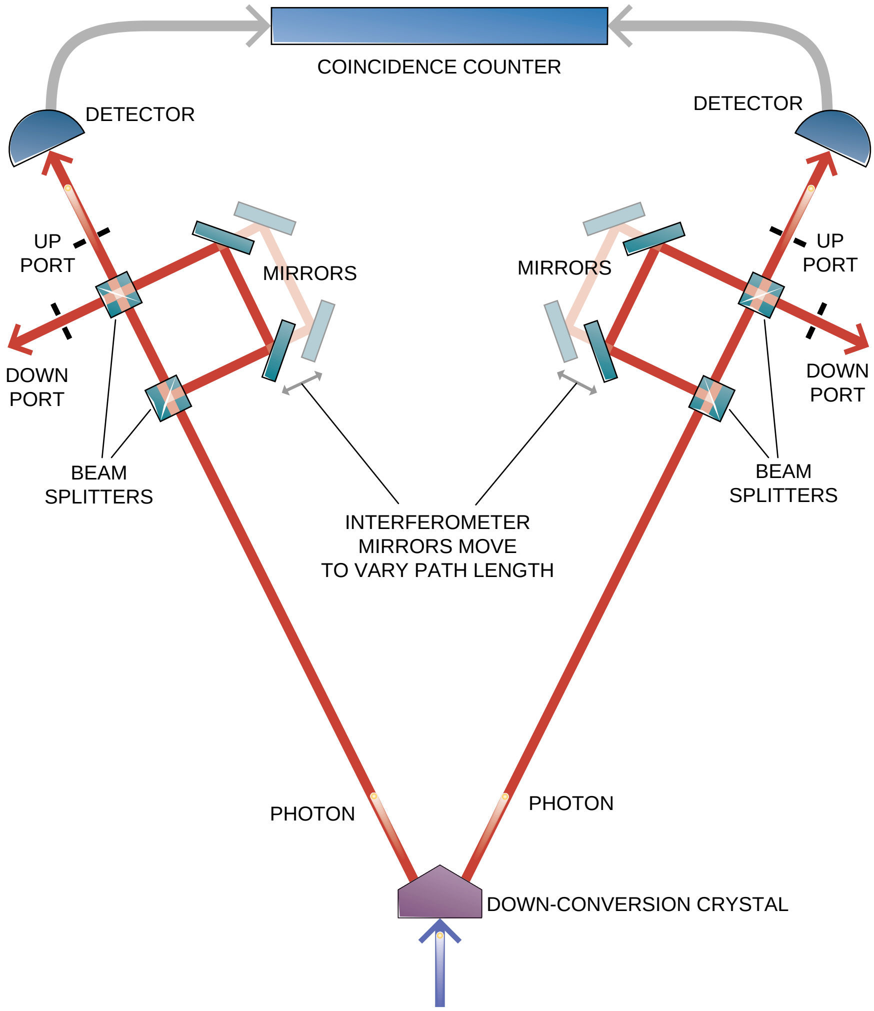
Now watch what happens when we send the members of a pair of photons through these interferometers. Each photon will randomly choose the long route (through the detour) or the shorter, direct route. After following one of the two paths, a photon can leave its interferometer through either of two ports, one labeled "up" and the other "down." We observed that each particle was as likely to leave through the up port as it was through the down. Thus, one might intuitively presume that the photon's choice of one exit would be unrelated to the exit choice its twin makes in the other interferometer. Wrong. Instead we see strong correlations between which way each photon goes when it leaves its interferometer. For certain detour lengths, for example, whenever the photon on the left leaves at the up exit, its twin on the right flies through its own up exit.

One might suspect that this correlation is built in from the start, as when one hides a white pawn in one fist and a black pawn in the other. Because their colors are well defined at the outset, we are not surprised that the instant we find a white pawn in one hand, we know with certainty that the other must be black.

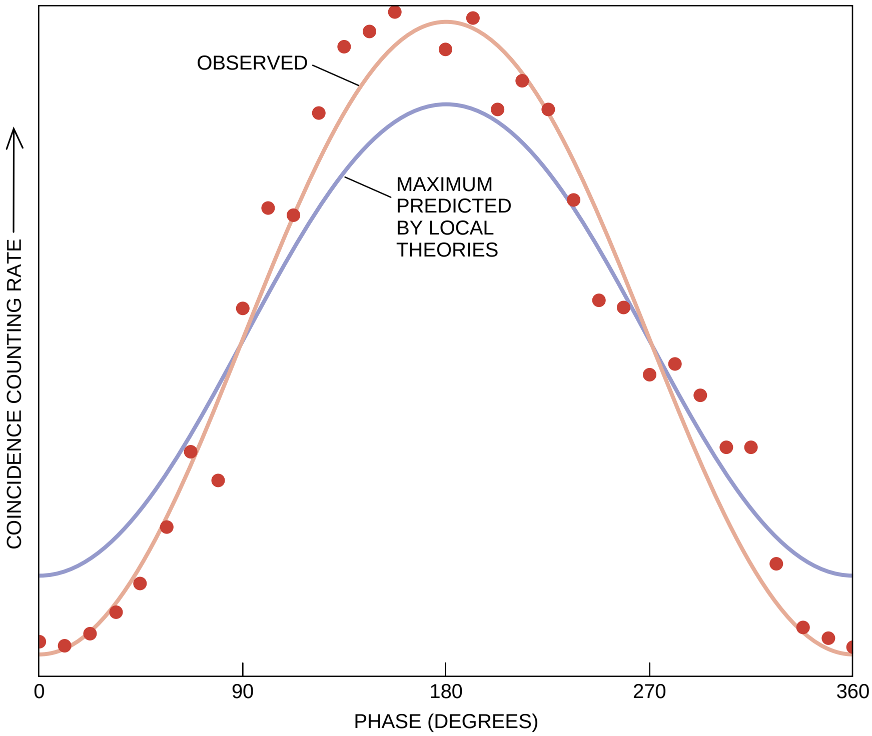
But a built-in correlation cannot account for the actual case in our experiment, which is much stranger: by changing the path length in either interferometer, we can control the nature of the correlations. We can go smoothly from a situation where the photons always exit the corresponding ports (both use the up port, or both use the down port) of their respective interferometers to one in which they always exit opposite ports. In principle, such a correlation would exist even if we adjusted the path length after the photons had left the source. In other words, before entering the interferometer, neither photon knows which way it is going to have to go—but on leaving, each one knows instantly (nonlocally) what its twin has done and behaves accordingly.

To analyze these correlations, we look at how often the photons emerge from each interferometer at the same time and yield a coincidence count between detectors placed at the up exit ports of the two interferometers. Varying either of the long-arm path lengths does not change the rate of detections at either detector individually. It does, however, affect the rate of coincidence counts, indicating the correlated behavior of each photon pair. This variation produces "fringes" reminiscent of the light and dark stripes in the traditional two-slit interferometer showing the wave nature of particles.

In our experiment, the fringes imply a peculiar interference effect. As alluded to earlier, interference can be expressed as the result of two or more indistinguishable, coexisting possibilities leading to the same final outcome (recall our second example of nonlocality, in which each photon travels along two different paths simultaneously, producing interference). In the present case, there are two possible ways for a coincidence count to occur: either both photons had to travel the short paths, or both photons had to travel the long paths. (In the cases in which one photon travels a short path and the other a long path, they arrive at different times and



NONLOCAL CORRELATION between two particles is demonstrated in the so-called Franson experiment, which sends two photons to separate but identical interferometers. Each photon may take a short route or a longer "detour" at the first beam splitter. They may leave through the upper or lower exit ports. A detector looks at the photons leaving the upper exit ports. Before entering its interferometer, neither photon knows which way it will go. After leaving, each knows instantly and nonlocally what its twin has done and so behaves accordingly.



RATE OF COINCIDENCES between left and right detectors in the Franson experiment (*red dots, with best-fit line*) strongly suggests nonlocality. The horizontal axis represents the sum of the two long path lengths, in angular units known as phases. The "contrast," or the degree of variation in these rates, exceeds the maximum allowed by local, realistic theories (*blue line*), implying that the correlations must be nonlocal, as shown by John S. Bell of CERN.

so do not interfere with each other; we discard these counts electronically.)

The coexistence of these two possibilities suggests a classically nonsensical picture. Because each photon arrives at the detector at the same time after having traveled both the long and short routes, each photon was emitted "twice"—once for the short path and once for the long path.

To see this, consider the analogy in which you play the role of one of the detectors. You receive a letter from a friend on another continent. You know the letter arrived via either an airplane or a boat, implying that it was mailed a week ago (by plane) or a month ago (by boat). For an interference effect to exist, the one letter had to have been mailed at both times. Classically, of course, this possibility is absurd. But in our experiments the observation of interference fringes implies that each of the twin photons possessed two indistinguishable times of emission from the crystal. Each photon has two birthdays.

More important, the exact form of the interference fringes can be used to differentiate between quantum mechanics and any conceivable local hidden variable theory (in which, for example, each photon might be born with a definite energy or already knowing which exit port to take). According to the constraints derived by Bell, no hidden variable theory can predict sinusoidal fringes that exhibit a "contrast" of greater than 71 percent—that is, the difference in intensity between light and dark stripes has a specific limit. Our data, however, display fringes that have a contrast of about 90 percent. If certain reasonable supplementary assumptions are made, one can conclude from these data that the intuitive, local, realistic picture suggested by Einstein and his cohorts is wrong: it is impossible to explain the observed results without acknowledging that the outcome of a measurement on the one side depends nonlocally on the result of a measurement on the other side.

So is Einstein's theory of relativity in danger? Astonishingly, no, because there is no way to use the correlations between particles to send a signal faster than light. The reason is that whether each photon reaches its detector or instead uses the down exit port is a random result. Only by comparing the apparently random records of counts at the two detectors, necessarily bringing our data together, can we notice the nonlocal correlations. The principles of causality remain inviolate.

Science-fiction buffs may be saddened to learn that faster-than-light communication still seems impossible. But several scientists have tried to make the best of the situation. They propose to use the randomness of the correlations for various cipher schemes. Codes produced by such quantum cryptography systems would be absolutely unbreakable [see "Quantum Cryptography," by Charles H. Bennett, Gilles Brassard and Artur D. Ekert; SCIENTIFIC AMERICAN, October 1992].

We have thus seen nonlocality in three different instances. First, in the process of tunneling, a photon is able to somehow sense the far side of a barrier and cross it in the same amount of time no matter how thick the barrier may be. Second, in the high-resolution timing experiments, the cancellation of dispersion depends on each of the two photons having traveled both paths in the interferometer. Finally, in the last experiment discussed, a nonlocal correlation of the energy and time between two photons is evidenced by the photons' coupled behavior after leaving the interferometers. Although in our experiments the photons were separated by only a few feet, quantum mechanics predicts that the correlations would have been observed no matter how far apart the two interferometers were.

Somehow nature has been clever enough to avoid any contradiction with the notion of causality. For in no way is it possible to use any of the above effects to send signals faster than the speed of light. The tenuous coexistence of relativity, which is local, and quantum mechanics, which is nonlocal, has weathered yet another storm.

FURTHER READING

QED: THE STRANGE THEORY OF LIGHT AND MATTER. Richard P. Feynman. Princeton University Press, 1985.

SPEAKABLE AND UNSPEAKABLE IN QUANTUM MECHANICS. J. S. Bell. Cambridge University Press, 1988.

HIGH-VISIBILITY INTERFERENCE IN A BELL-INEQUALITY EXPERIMENT FOR ENERGY AND TIME. P. G. Kwiat, A. M. Steinberg and R. Y. Chiao in *Physical Review A*, Vol. 47, No. 4, pages R2472–R2475; April 1, 1993.

THE SINGLE-PHOTON TUNNELING TIME. A. M. Steinberg, P. G. Kwiat and R. Y. Chiao in *Proceedings of the XXVIIIth Rencontre de Moriond*. Edited by Jean Trân Thanh Vân. Editions Frontières, Gif-sur-Yvette, France (in press).

SUPERLUMINAL (BUT CAUSAL) PROPAGATION OF WAVE PACKETS IN TRANSPARENT MEDIA WITH INVERTED ATOMIC POPULATIONS. Raymond Y. Chiao in *Physical Review A* (in press).

T Cell Anergy

When cells of the immune system "see" antigens in the absence of the right cosignals, they shut themselves down instead of attacking. Future therapies might capitalize on that reaction

by Ronald H. Schwartz

The immune system contains billions of white blood cells called lymphocytes. These cells prowl the body, looking for evidence of foreign or unfamiliar materials that might indicate a virus, bacterium, parasite or tumor has infiltrated the tissues. Once an intruder has been detected, the lymphocytes mount a defense against it. To optimize its detection capabilities, the immune system has evolved a genetic mechanism for placing a unique molecular receptor on the surface of each lymphocyte. The diversity of these receptors enables the immune system to recognize an almost unlimited array of foreign molecules. Yet at the same time, this mechanism creates a problem for the individual: some of the molecules to which the system can react are constituents of the body's normal tissues. Because such responses would be harmful, the immune system must go through a series of modifications that prevent self-destruction. This process is known as tolerance induction.

Immunologists have identified several mechanisms that induce tolerance. First and foremost is the physical elimination of lymphocytes that recognize the body's molecules during their development, which is referred to as clonal deletion. A second process, immunoregulation, involves the generation of regulatory cells that weaken harmful or inappropriate lymphocyte responses. Recently a third mechanism—anergy—has become apparent. Anergy turns off self-reactive cells as a consequence of the way in which the molecules are presented to them.

Although detailed study of anergy has only just begun, biologists have already identified some of the molecules that seem to mediate it. No one yet knows whether anergy plays a major or minor role in producing tolerance. Nevertheless, many researchers are hopeful that the power of anergy to check immune responses may yet be harnessed therapeutically. The selective induction of anergy could prove extraordinarily useful in preventing the rejection of transplanted organs. It also holds promise for treating patients afflicted with autoimmune diseases, who carry lymphocytes that are misguidedly attacking some of their tissues.

To understand anergy, one must look at how the immune system identifies and responds to pathogens. The first step involves the preparation of a molecule of foreign protein, referred to as an antigen, for identification. Interdigitating dendritic cells and other so-called antigen-presenting cells scavenge materials from the blood and tissues and digest them. Inside the cells, small pieces of the antigens then bind to a special class of proteins called major histocompatibility complex (MHC) molecules.

The MHC molecules transport the antigen fragments to the surface of the presenting cell and display them to passing lymphocytes known as helper *T* cells, which coordinate and execute many disease-fighting functions. Each *T* cell is equipped with antigen-specific receptor molecules that enable it to recognize just one type of antigen fragment on an MHC molecule. If a *T* cell finds a matching antigen on a presenting cell and if that presenting cell offers the appropriate signals, the lymphocyte responds in two major ways. One is to enlarge and divide, thereby increasing the number of cells that react to the antigen. The other is to secrete lymphokines, proteins that directly inhibit the pathogen or that recruit other cells to join in the immune response.

For example, helper *T* cells make lymphokines called interleukins that stimulate the other major population of lymphocytes, *B* cells, to make antibodies. Antibodies bind selectively to antigens; in fact, they are soluble forms of the antigen receptor molecules on the *B* cells. Yet the binding of an antigen to a *B* cell's receptor is not sufficient for activation: the *B* cell needs help from a *T* cell. Toward that end, the *B* cell internalizes the antigen, digests it as a presenting cell would and displays the pieces on its surface bound to MHC molecules. If a *T* cell recognizes the displayed antigen, it becomes active and produces lymphokines and other molecules. Those chemicals, in turn, complete the stimulation of the *B* cell. As a result, antibodies are released into the circulation to find and neutralize infectious agents throughout the body. Lymphokines from helper *T* cells also activate another subpopulation of lymphocytes, cytotoxic *T* cells, which kill cells in tissues infected with viruses and bacteria. These antibody and *T* cell responses form two prongs of the immunologic defense against pathogens.

Like *B* cells and cytotoxic *T* cells, the helper *T* cells need more than the signal that comes from recognizing an antigen to become activated. As diverse evidence shows, they require a second molecular signal from the antigen-presenting cell. The earliest experiment

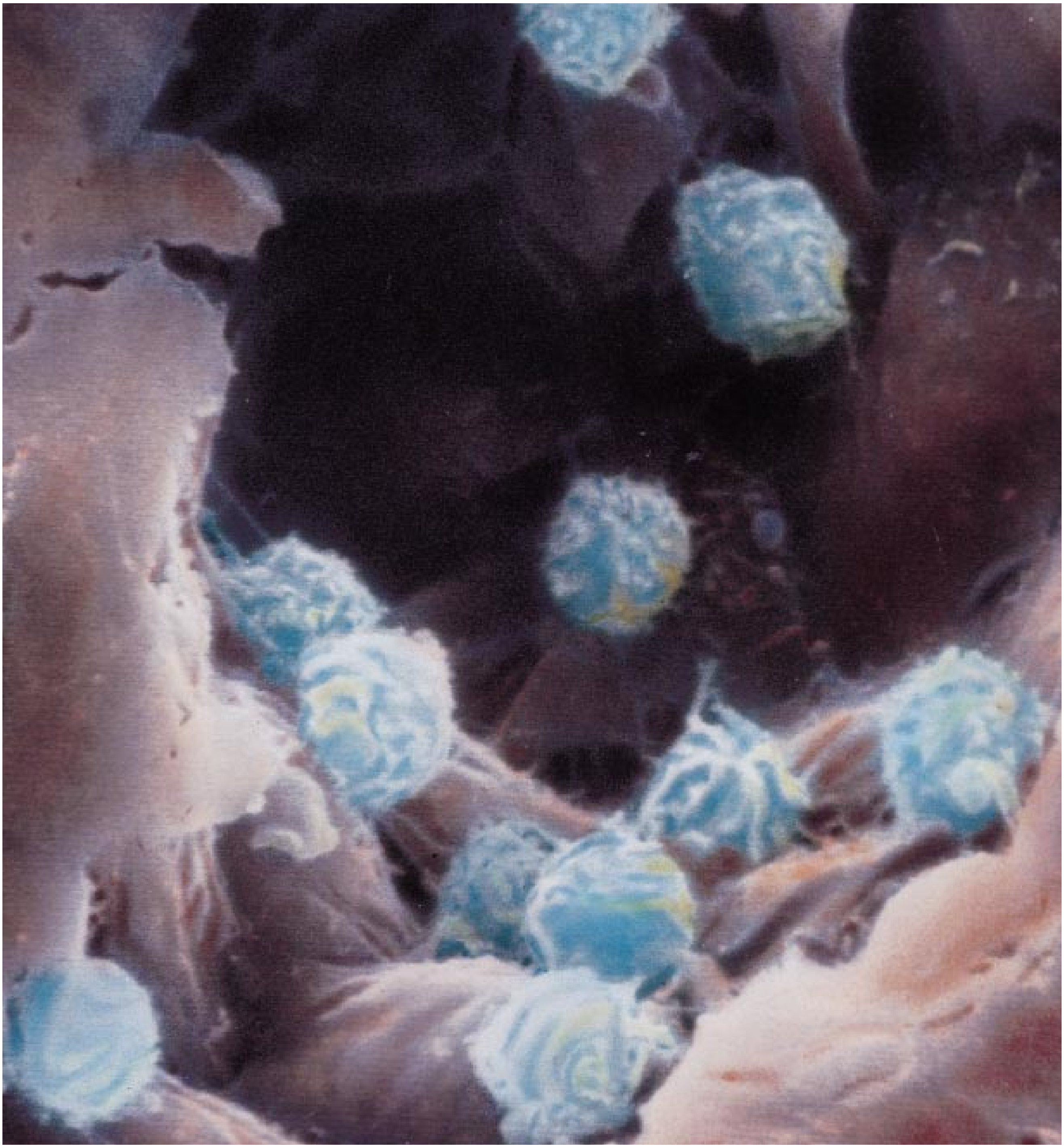
_T_ LYMPHOCYTES are white blood cells that regulate much of the immune system's response to invading microorganisms and foreign materials in the body. Researchers have discovered that a process called anergy can "turn off" specific sets of _T_ cells. Anergy may partially explain why the immune system does not ordinarily attack healthy tissues.

RONALD H. SCHWARTZ, one of the discoverers of *T* cell anergy, is the chief of the Laboratory of Cellular and Molecular Immunology at the National Institute of Allergy and Infectious Diseases. After graduating from Cornell University in 1965 with a degree in chemistry, he went on to attain an M.D. from Harvard Medical School in 1970 and a doctorate in microbiology from Rutgers University in 1973. Schwartz is a reviewing editor for several immunology journals and an adviser to the biotechnology industry. In 1991 he received the U.S. Public Health Service Distinguished Service Medal.

suggesting that two signals are necessary was performed in 1962 by David W. Dresser while at the California Institute of Technology. He found that if he removed all the aggregates, or clumps of protein, from samples of foreign antigens before injecting them into mice, the remaining, soluble proteins failed to elicit a normal immune response, as measured by the production of antibodies. The soluble proteins became immunogenic again only after Dresser emulsified them in an adjuvant, a compound developed by the late Jules Freund of the National Institute of Allergy and Infectious Diseases (NIAID). This adjuvant was a "witches' brew" composed of mineral oil, saline solution, a detergent and a heat-killed form of the organism that causes tuberculosis.

The most intriguing finding of Dresser's experiment was that the mice injected only with the deaggregated antigens became tolerant—that is, they failed to give a normal immune response even if they were subsequently challenged with the intact antigens or antigens in adjuvants. Dresser's interpretation of the results was that an antigen had to have two properties—foreignness and adjuvanticity—to be immunogenic. Foreignness alone, he surmised, paralyzed the immune system and inhibited future responses.

P. C. Frei, Baruj Benacerraf and G. Jeanette Thorbecke of New York University extended the theory. They interpreted the adjuvanticity property as a need for the antigen to be taken up by certain cells in the immune system, such as macrophages, for a response to be initiated. (Freund's adjuvant and the



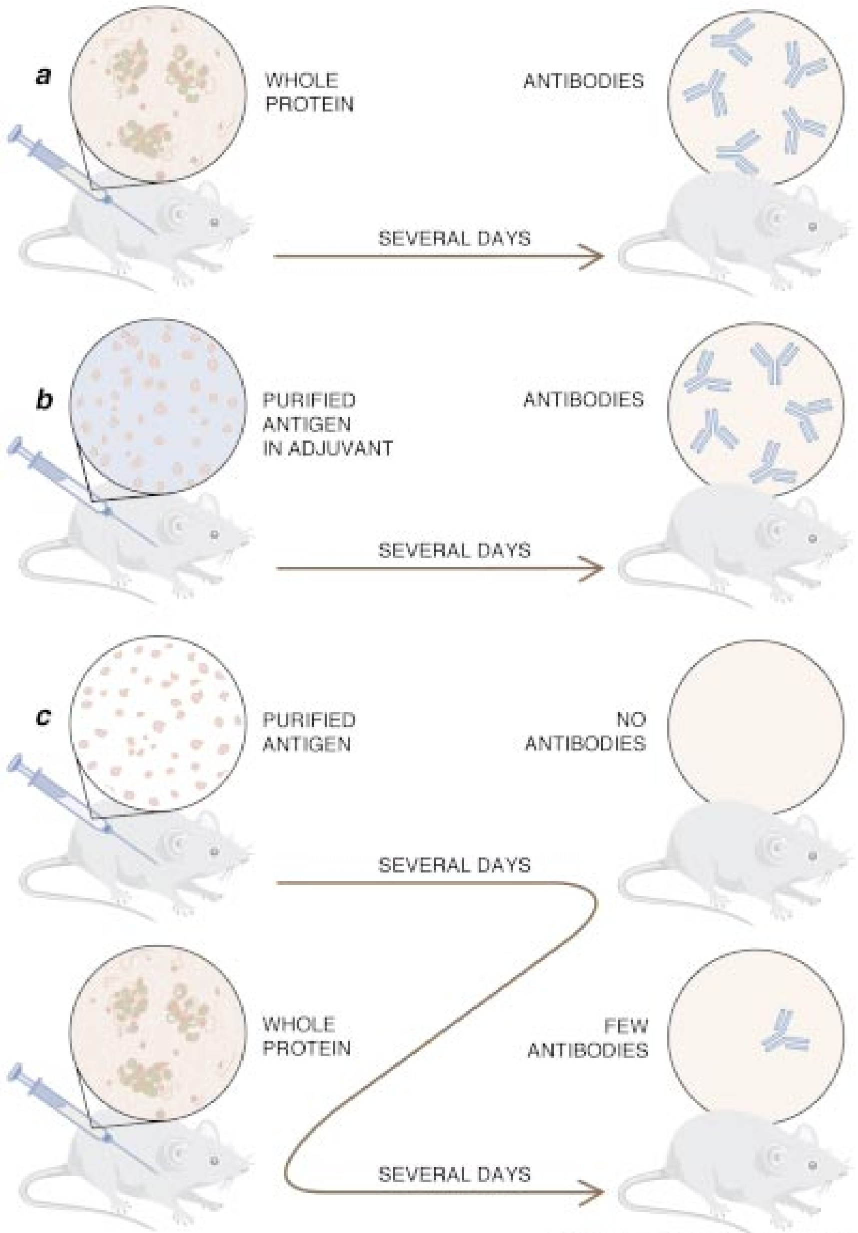
aggregates presumably made the antigen a more obvious or attractive target for those cells.) At that time, it was known that chemical extracts from macrophages that had been fed antigens were more immunogenic than the antigens themselves. The immunologists postulated that in Dresser's experiment, the purified antigens had bypassed the macrophages and gone directly to the lymphocytes and consequently lacked a macrophage component that was crucial for a positive response.

Two theoretical frameworks for understanding Dresser's observations were developed a few years later, by Peter A. Bretscher and Melvin Cohn of the Salk Institute for Biological Studies in San Diego in 1969 and by Kevin J. Lafferty and Alistair J. Cunningham of the John Curtin School of Medical Research in Canberra in 1974. The essence of the Bretscher and Cohn model is that two signals are required to turn on a resting lymphocyte and that receiving the first signal alone will turn it off. In their model, both signals are antigen-specific: the lymphocyte's recognition of the foreign substance generates both signals separately. The Lafferty and Cunningham model also rests on the need for two signals, but it proposes that the second signal does not depend on recognition of the antigen—a notion that fits better with Dresser's nonspecific adjuvant results. The second signal has no effect on its own, but in conjunction with the antigen-specific signal, it will activate a lymphocyte.

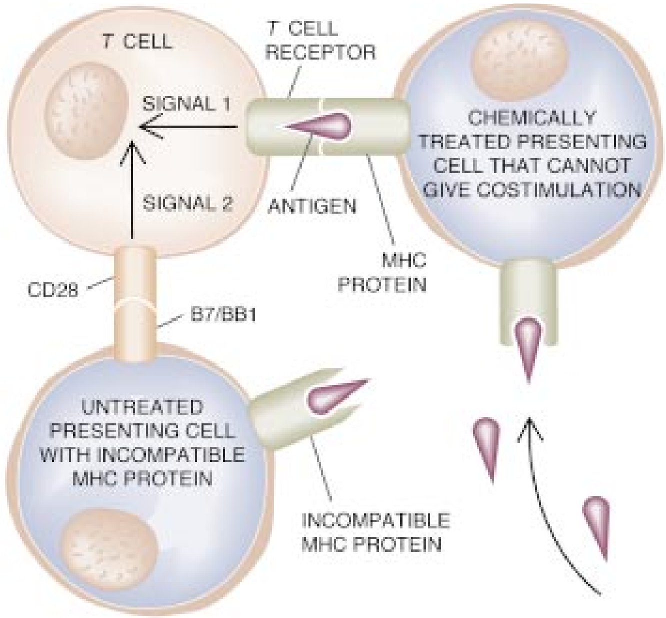
Unfortunately, further work on the problem of lymphocyte activation stalled until the 1980s, when experimental immunology had at last advanced far enough technically to test these models. Of particular importance was the development of procedures for cloning *T* lymphocytes, which made it possible to produce large numbers of antigen-specific cells in tissue culture for analyses. Researchers saw that *T* cells could be stimulated to divide and to make lymphokines when presented with their specific antigens. Antigen alone, however, was not sufficient: antigen-presenting cells also had to participate. It eventually became clear that the additional cells processed and presented the antigen bound to MHC molecules and that the *T* cell receptor could not recognize an antigen in the absence of an MHC molecule.

Marc Jenkins and I at NIAID were the first to show that antigen-presenting cells also deliver the critical second signal. In the test tube, using chemicals such as carbodiimides and paraformaldehyde, we stopped presenting cells from processing antigens. Because these cells still carried MHC molecules on their surface, they were able to bind antigen fragments that we had treated chemically and applied to the cells externally. The inactivated cells could therefore still present antigens to *T* lymphocytes. Nevertheless, they did not activate clones of *T* cells until we added another set of presenting cells that carried a different MHC molecule. Because their MHC molecules could not bind to the relevant antigen, they could not have given the *T* cells an antigen-related signal. Instead they were providing a second, nonspecific signal.

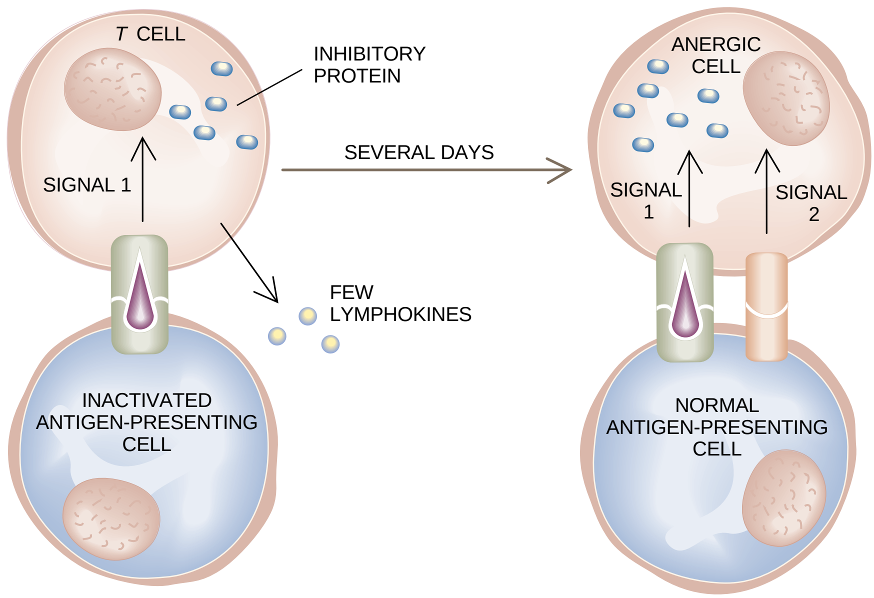
This three-cell experimental model demonstrated that the second signal could be delivered independently of the antigen-specific signal. (Normally, of course, both signals would be presented by the same cell.) In molecular terms, that result implied that a second signaling system, involving another re-



MORE THAN ANTIGEN is needed to elicit an effective immune response. If mice are injected with a whole protein preparation, they develop antibodies against that antigen (*a*). Similarly, the mice also develop antibodies if they are injected with purified proteins in an emulsion of adjuvant (*b*). If the mice receive only the purified antigen, however, they do not respond (*c*). Moreover, those same mice will later evince no more than a weak response even if they receive injections of whole protein. Exposure to the purified antigen alone has rendered the mice tolerant.



THREE-CELL EXPERIMENT demonstrated that the two signals essential for *T* cell activation could be delivered separately. One signal was given by chemically inactivated cells presenting antigens that had been added externally. The second signal came from cells that did not display an antigen to which the *T* cells reacted. That signal results from an interaction between B7/BB1 proteins on the presenting cells and a CD28 receptor molecule on the lymphocytes.



ANERGY INDUCTION occurs when a *T* cell is presented with the antigen-specific signal but without the cosignal. The *T* cell produces low levels of some lymphokines (but not interleukin-2) and also an inhibitory protein. Later, even if the anergic *T* cell does receive both signals, it does not divide.

ceptor on the *T* cell and another molecule on the antigen-presenting cell, mediated the event. In recent years, several laboratories have published evidence that suggests the second *T* cell receptor is a protein called CD28. A related receptor—CTLA-4—is also expressed when the *T* lymphocyte is activated. The molecule to which the receptors bind is B7/BB1, a protein that appears on the surface of presenting cells when they are activated. The presence of B7/BB1 on a cell may be induced as a consequence of the interaction of the antigen-specific *T* cell receptor with the MHC molecule on the presenting cell. It is also induced by other stimuli, such as bacterial substances, that might be considered analogous to those delivered by adjuvants. Our experiment thus verified the existence of all the elements essential to the two-signal model of Lafferty and Cunningham.

It also confirmed the most significant part of the Bretscher and Cohn model—that by itself the signal associated with antigen recognition would have a negative consequence for a *T* cell. The *T* cells did appear to recognize and respond to the antigens displayed by chemically treated presenting cells: they increased in size and released small amounts of certain lymphokines. Yet they did not make the lymphokine most critical for *T* cell division, interleukin-2. Moreover, when attempts were made several days later to restimulate the *T* lymphocytes using normal presenting cells that could deliver both the antigen-specific and the second, costimulatory signals, the *T* cells still did not divide or produce much interleukin-2, and their production of other lymphokines was reduced to varying degrees.

This unresponsive state is generally referred to as *T* cell anergy. The term "anergy"—"without working"—was borrowed from Sir Gustav Nossal of the Walter and Eliza Hall Institute of Medical Research in Melbourne, who had previously coined it to describe a possibly analogous inactive state in *B* cells. We and other researchers observed that in culture, clones of *T* cells could remain alive in the anergic state for weeks, although recent quantitative studies suggest that the state slowly decays. Anergy can also be reversed rapidly if interleukin-2 is given to the *T* cells to make them divide.

How do the stimulation of interleukin-2 production and the induction of anergy come about at the molecular level? As Helen Quill in my laboratory initially demonstrated, all that is required to deliver the first signal is an appropriate antigen bound to the MHC molecule. She analyzed the sequence of chemical events involved in the transduction of this signal into *T* cells using a lipid membrane system developed by Harden M. McConnell of Stanford University and his associates. First, the antigen-specific receptor on a *T* cell recognizes an antigen-MHC complex and sends signals into the cell. The immediate effect, as shown by Lawrence E. Samelson of the National Institute of Child Health and Human Development and Arthur Weiss of the University of California at San Francisco, is the stimulation of a tyrosine kinase, an enzyme that adds energy-rich phosphate groups to the amino acid tyrosine in proteins.

That event begins a chain of chemical reactions. By adding a phosphate, the tyrosine kinase activates the enzyme phospholipase C γ1, which in turn cleaves a compound in the cell membrane called phosphatidylinositol bisphosphate. One of these cleavage products activates another enzyme, protein kinase C, which adds phosphate groups to the amino acids serine and threonine on other proteins. Through a series of reactions that have not yet been entirely worked out, the activities of protein kinase C eventually lead to the synthesis of a protein complex called AP-1.

Meanwhile the other cleavage product triggers an influx of calcium ions into the cell, which leads to further enzyme activity. As a consequence, a molecule known as the nuclear factor of activated *T* cells (NF-AT) is modified in such a way that it can migrate from the cytoplasm into the nucleus. Once there, the NF-AT binds to the newly synthesized AP-1. The two proteins form a complex that then binds tightly to the DNA located next to particular genes. Such nuclear protein complexes are referred to as transcription factors because their binding to the DNA allows the gene to be transcribed, or copied, into the form of messenger RNA molecules, which the cellular machinery later translates into new proteins.

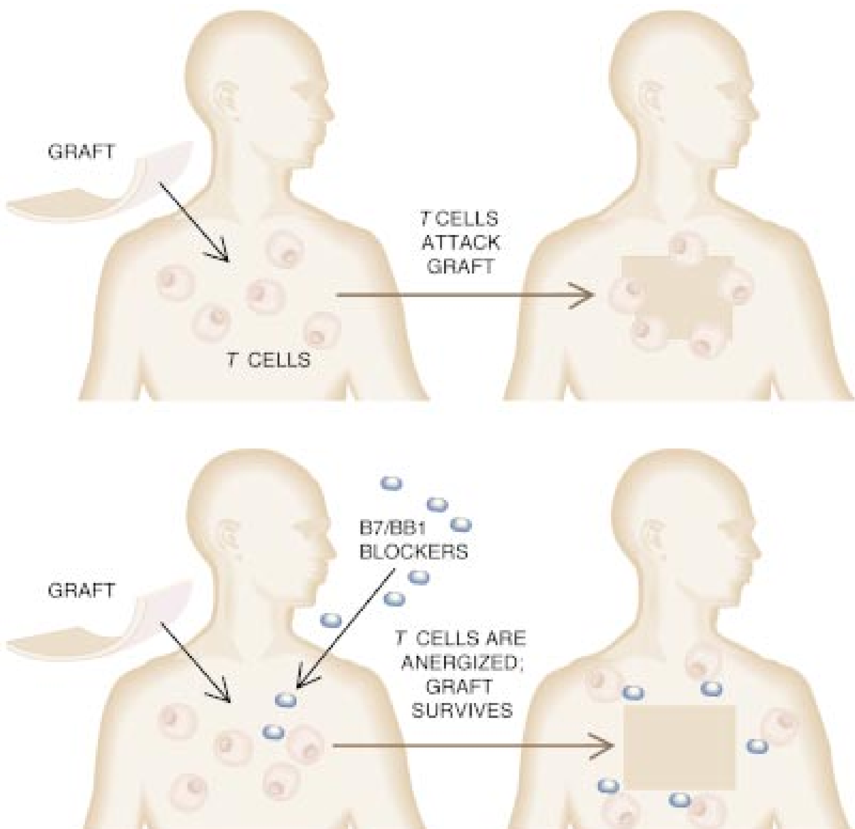
Many lymphokine genes can be transcribed if only a few transcription factors bind to their control regions. The gene for interleukin-2, however, seems to be under very tight regulation: it cannot be transcribed unless at least the NF-AT/AP-1 complex, AP-1 by itself and two other transcription factors are present. All the factors must be newly made or chemically activated when the *T* cell is stimulated through its antigen-specific receptor. Moreover, even the presence of all four transcription factors is not adequate to allow the cell to produce significant quantities of interleukin-2. A stimulus from the second signal, which is mediated by the CD28 molecule, is also required. The biochemical pathway for this signaling event is not known, although recent work from the laboratory of Carl H. June of the Naval Medical Research Institute suggests that the first step might involve the activation of another tyrosine kinase.

Experiments have raised two possible mechanisms for how costimulation could amplify interleukin-2 production. One, from Weiss and his colleagues, is by activating yet another transcription factor (CD28RC), which provides the final component required for messenger RNA production. The other, from the laboratory of Tullia Lindsten of the University of Michigan, is by stabilizing the messenger RNA transcribed from the gene. In the cytoplasm the interleukin-2 messenger RNA can degrade rapidly. Any modification of the machinery that performs this degradation would enable the messenger RNA to exist longer, which would increase the opportunity for the message to be translated into interleukin-2. Whichever mechanism is at work, however, a cell cannot produce enough interleukin-2 for division without costimulation.

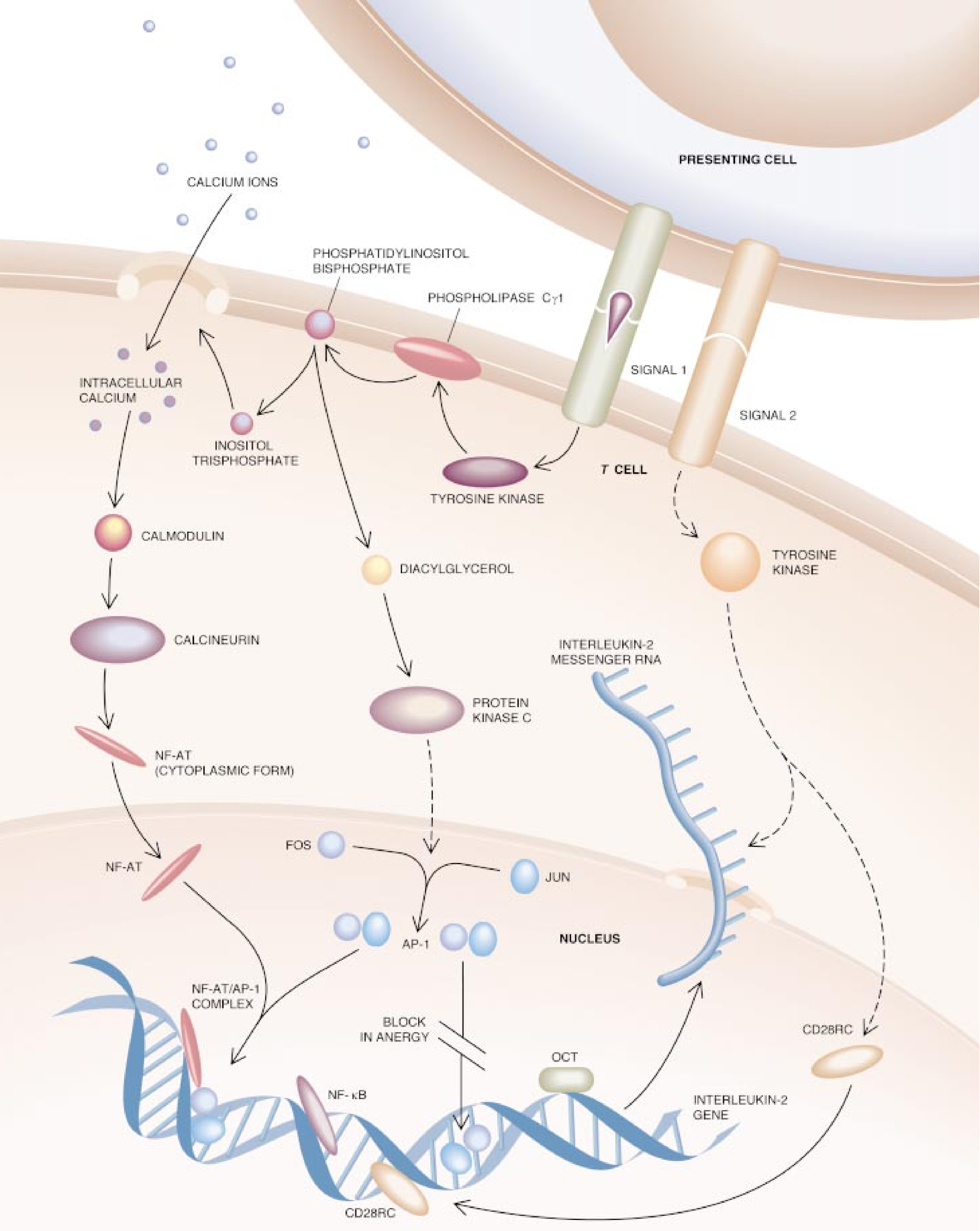
The other effect of the absence of costimulation is that the cell is driven into the anergic state. At this point, researchers can only speculate why. We do know that drugs that prevent protein synthesis also block the induction of anergy. The antigen signal, when presented by itself, may therefore lead to the production of an inhibitory protein. Recent studies done by our laboratory in collaboration with that of Michael J. Lenardo of NIAID and his colleagues were relevant to that idea. We showed that in anergic *T* cells, the binding of the AP-1 transcription factor to the regulatory region of the interleukin-2 gene is reduced. The inhibitory protein might accomplish that end in several ways. It might block the transcription of the genes that make the subunits of AP-1. It might bind with the interleukin-2 gene to prevent the attachment of the AP-1 transcription factor. The inhibitor might also chemically modify or bind to AP-1 in such a way that the AP-1 will no longer bind to the interleukin-2 gene. Any of those effects would have the net result of preventing the synthesis of interleukin-2.

How does the costimulatory signal prevent the induction of anergy? According to the current model, when both the antigen-specific and the costimulatory signals are presented to a *T* cell, it synthesizes interleukin-2. That interleukin-2 then binds to receptors on the *T* cell's surface and triggers cell division. Either the process of division itself or some other intracellular signals associated with the binding of interleukin-2 seems to block the action of the inhibitory protein. Although we do not yet understand the exact molecular mechanism by which this happens, one possibility is that the division of the cell dilutes the concentration of the inhibitory protein to an ineffective level.

One of the most pressing needs of current immunologic research is to find animal models for *T* cell anergy that can be used to determine whether the observations made in tissue culture



GRAFT REJECTION that can occur after transplant operations might be prevented by inducing *T* cell anergy. *T* cells will ordinarily attack and kill many grafts (*top*) unless the patient receives broad immunosuppressive drugs. If substances that blocked the attachment of B7/BB1 molecules to *T* cells were administered before or during the transplantation, the *T* cells would not receive the second activation signal (*bottom*). The cells would therefore become anergic instead of attacking.


PRESENTING CELL
CALCIUM IONS
PHOSPHATIDYLINOSITOL
BISPHOSPHATE
PHOSPHOLIPASE Cγ1
SIGNAL 1
SIGNAL 2
INTRACELLULAR
CALCIUM
INOSITOL
TRISPHOSPHATE
T CELL
TYROSINE KINASE
CALMODULIN
DIACYLGLYCEROL
TYROSINE
KINASE
CALCINEURIN
INTERLEUKIN-2
MESSENGER RNA
PROTEIN
KINASE C
NF-AT
(CYTOPLASMIC FORM)
FOS
NF-AT
JUN
AP-1
NUCLEUS
NF-AT/AP-1
COMPLEX
BLOCK
IN ANERGY
CD28RC
OCT
NF- κB
INTERLEUKIN-2
GENE
CD28RC

SIGNALING PATHWAYS involved in *T* cell responses and anergy have been partially determined. When a *T* cell receives the antigen-specific signal, a tyrosine kinase enzyme is activated, initiating a cascade of other chemical changes. One consequence of these events is that the Fos and Jun components of the AP-1 protein are synthesized. Another effect is that the protein NF-AT can enter the cell's nucleus, form a complex with AP-1 and bind to the interleukin-2 gene. NF-AT/AP-1 by itself and the factors OCT and NF-κB are all essential for the transcription of the gene. The second signal seems to boost the production of interleukin-2 to effective levels. It may do this by producing the transcription factor CD28RC or by somehow stabilizing the messenger RNA transcribed from the interleukin-2 gene. In anergy, an inhibitory protein may block the attachment of AP-1 to the interleukin-2 gene.

are relevant to normal lymphocyte biology. In particular, immunologists must learn whether anergy is solely a feedback mechanism for restraining cells after an immune response is under way or rather an important mechanism of tolerance induction.

Susumu Tonegawa of the Massachusetts Institute of Technology and his colleagues have created one good animal model that has offered some insights. They genetically engineered a mouse in which almost all the lymphocytes expressed the same antigen receptor. They then mated it with a mouse expressing that antigen to produce offspring whose *T* cells would react to their own tissues. Yet as the researchers discovered when they looked in the spleens of the offspring mice, all the *T* cells present were anergic. Their anergy appeared to be similar to that observed in cultured *T* cell clones: the cells failed to make interleukin-2 when stimulated with antigen but did proliferate if they were also given interleukin-2. Those results suggest that *T* cell anergy is a part of the way in which immune systems become tolerant of self-molecules.

How might tolerance through anergy arise inside the body? Nearly all the cells in the body have MHC molecules; the difference is that specialized presenting cells can deliver costimulation, whereas other cells cannot. As *T* cells roam through tissues, they examine the MHC molecules and antigens they find there. Occasionally, a *T* cell may encounter a self-antigen on a body cell to which it can respond. If that cell cannot supply the essential second signal—that is, if it cannot express the B7/BB1 molecule—the lymphocyte will receive only one signal and will fall into an anergic state rather than an immunologically active one.

Regardless of whether anergy turns out to be a major mechanism of tolerance induction, knowledge of the molecular details of how to bring it about should have important implications for immunotherapy. In transplantation, donor organs bearing MHC molecules of one type are often given to recipients with different MHC molecules. Barring intervention, that incompatibility would cause the recipient's immune system to reject the graft. Currently physicians prevent the rejection by administering immunosuppressive drugs such as cyclosporine, which shut down most immune responses, thereby leaving the patient potentially vulnerable to serious infection. Ideally, one would like to suppress only the specific immune response against the donor organ.

Anergy induction presents a plausible means for accomplishing that end. During the transplant operation, the surgical team might use drugs that prevented costimulation—for example, the drugs might interfere with the expression of B7/BB1, the stimulating molecule, on the antigen-presenting cells or block the CD28 receptor on *T* cells. The *T* cells in the patient that would ordinarily attack the graft would consequently receive only the antigen recognition signal and become anergized; other *T* cells in the body would be unaffected.

Recently Jeffrey A. Bluestone of the University of Chicago and his colleagues succeeded in preventing mice from rejecting grafts of human pancreatic islet cells. They accomplished that end using a soluble form of the CTLA-4 receptor, engineered by Peter S. Linsley and others at Bristol-Meyers Squibb, which can block the B7/BB1 molecules presented on cells. The experience also induced tolerance in the mice: the animals accepted a second graft from the same human donor without further treatment. On the other hand, when Craig B. Thompson of the University of Michigan and his co-workers tried similar experiments with heart grafts in rats, the treatments greatly extended the survival of the grafts but did not prevent their ultimate rejection.

At this point, it is not clear whether soluble CTLA-4 can totally prevent costimulation. Furthermore, new *T* cells emerge from the thymus gland every day, and they may be capable of rejecting the graft until all the specialized presenting cells have migrated out of the tissue. It is conceivable, then, that a more prolonged administration of soluble CTLA-4 and the addition of other blocking agents will be required for human patients to benefit fully from anergy induction.

The value of anergy-based therapies would not be limited to transplant operations. Similar approaches should be applicable to the treatment of autoimmune diseases, such as diabetes mellitus, multiple sclerosis and rheumatoid arthritis. True, in the case of autoimmunity, the unwanted response is already occurring, whereas the treatment for graft rejection would be entirely preventive. Nevertheless, the work on anergy in cultured *T* cell clones shows that it is possible to induce activated lymphocytes to become unresponsive. Stephen D. Miller's laboratory at Northwestern University has studied an animal model of multiple sclerosis called experimental allergic encephalomyelitis. His group has induced tolerance to neurological antigens in rodents by chemically coupling the antigens to presenting cells and injecting them intravenously. That procedure stopped both the acute and the relapsing form of the disease. Although the precise mechanisms operating to induce tolerance in these experiments have not been fully elucidated, the method and the observation that the production of interleukin-2 diminished significantly imply strongly that anergy is playing a role.

Further work on *T* cell anergy will eventually trace the complete biochemical pathways that induce and maintain this unresponsive state. I am optimistic that one day we will be able to turn off specific subpopulations of *T* cells routinely and that this knowledge will lead to new and effective therapies in clinical medicine.

FURTHER READING

SPECIFIC INHIBITION OF ANTIBODY PRODUCTION II: PARALYSIS INDUCED IN ADULT MICE BY SMALL QUANTITIES OF PROTEIN ANTIGEN. David W. Dresser in *Immunology,* Vol. 5, No. 3, pages 378–388; March 1962.

A CELL CULTURE MODEL FOR T LYMPHOCYTE CLONAL ANERGY. Ronald H. Schwartz in *Science,* Vol. 248, pages 1349–1356; June 15, 1990.

THE ROLE OF CELL DIVISION IN THE INDUCTION OF CLONAL ANERGY. M. K. Jenkins in *Immunology Today,* Vol. 13, No. 2, pages 69–73; February 1992.

COSTIMULATION OF T LYMPHOCYTES: THE ROLE OF CD28, CTLA-4, AND B7/BB1 IN INTERLEUKIN-2 PRODUCTION AND IMMUNOTHERAPY. Ronald H. Schwartz in *Cell,* Vol. 71, No. 7, pages 1065–1068; December 24, 1992.

IMMUNOLOGIC TOLERANCE. Ronald H. Schwartz in *Fundamental Immunology,* third edition. Edited by William E. Paul. Raven Press (in press).

A Universe of Color

Color photography continues to be an important astronomical tool that reveals details of celestial objects not yet captured by modern electronic detectors

by David F. Malin

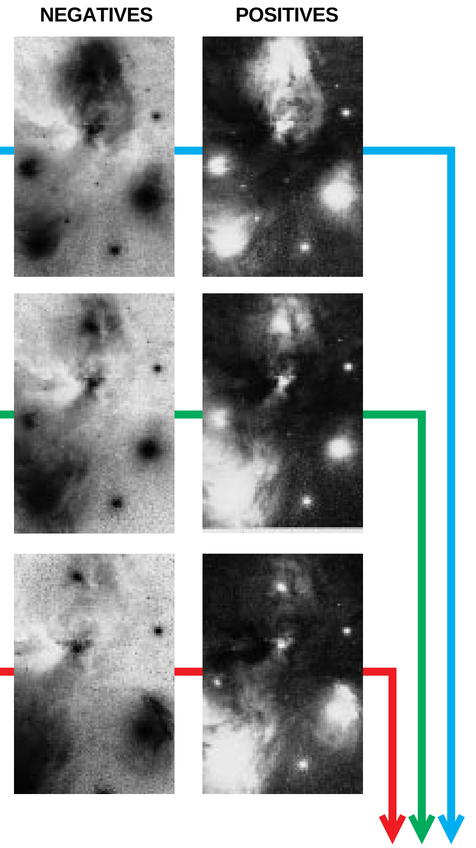
In the 150 years since the creation of the first crude daguerreotype of the moon, photography has evolved into an indispensable tool for astronomical research. That partnership has begun to dissolve during the past decade as most observers have switched to charge-coupled devices (CCDs), electronic detectors that collect light far more efficiently than do photographs. Nevertheless, the older process has attributes—especially the ability to record fine detail and accurate color over large areas—that CCDs cannot match. These assets are particularly evident in the stunning color photographs that my co-workers and I have made recently at the Anglo-Australian Observatory. Our approach brings to modern fruition a process the Scottish physicist James Clerk Maxwell developed more than a century ago.

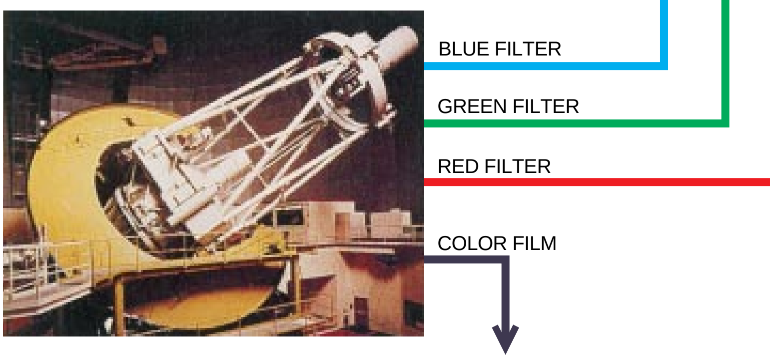
One of the key advantages photography has over electronic imaging is that photographic plates can capture high-resolution,

COMETARY GLOBULE is an isolated cloud of cold gas and dust whose tail consists of material swept out by radiation from nearby hot stars. The colors in this remarkable photograph speak eloquently about the physical nature of the globule known as CG4. The tail exhibits the blue tint typical of light reflected off fine dust particles. A surprising red glow around the head of the globule probably represents the characteristic emission from stimulated hydrogen atoms; the hydrogen was liberated from CG4 by energetic starlight. The blue portion of the nebula takes on a startling greenish hue where it is seen through foreground dust. This image contains some of the faintest details ever captured in a color photograph.

Recording the Colorful Cosmos

The three-color process depicted here can produce vastly superior results to those obtained using ordinary color film. A conventional color photograph of a nebula on the border between the constellations Scorpius and Ophiuchus, taken with the 1.2-meter U.K. Schmidt telescope (*below*), exhibits poor contrast and reveals little detail (*bottom left*). The author made three black-and-white negatives of the same nebula using combinations of filters and photographic emulsions that respond to blue, green or red light (*right*). These negatives were made into positives using a process that emphasizes the dimmest parts of the image. He then projected the positives, one at a time, through an appropriate color filter onto a piece of color negative film. The final, color positive (*bottom right*) reveals the nebula's elusive colors and intricate, wispy texture.



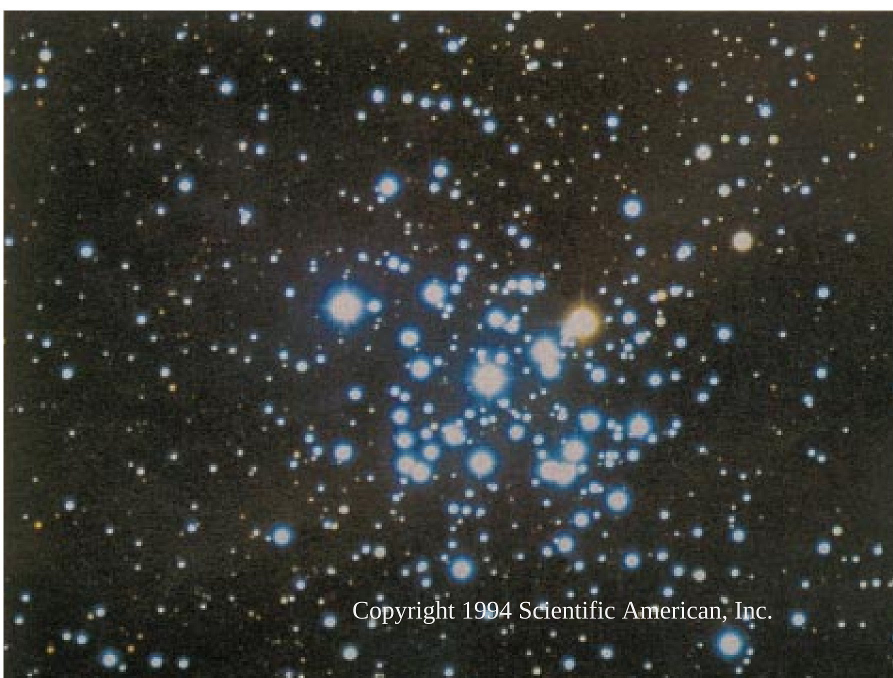


sensitive images over an area of virtually unlimited size, for example, across the entire wide field of view of the U.K. Schmidt telescope in Australia. In comparison, the largest CCDs measure only a few centimeters across. Photography therefore offers a superior means for recording images of extended astronomical objects such as nebulae and nearby galaxies.

Furthermore, the photographic layer serves as a compact medium for storing vast amounts of visual information. A single 35.6- by 35.6-centimeter plate from the U.K. Schmidt telescope, which records a square patch of sky 6.4 degrees on a side, contains the equivalent of several hundred megabytes of data.

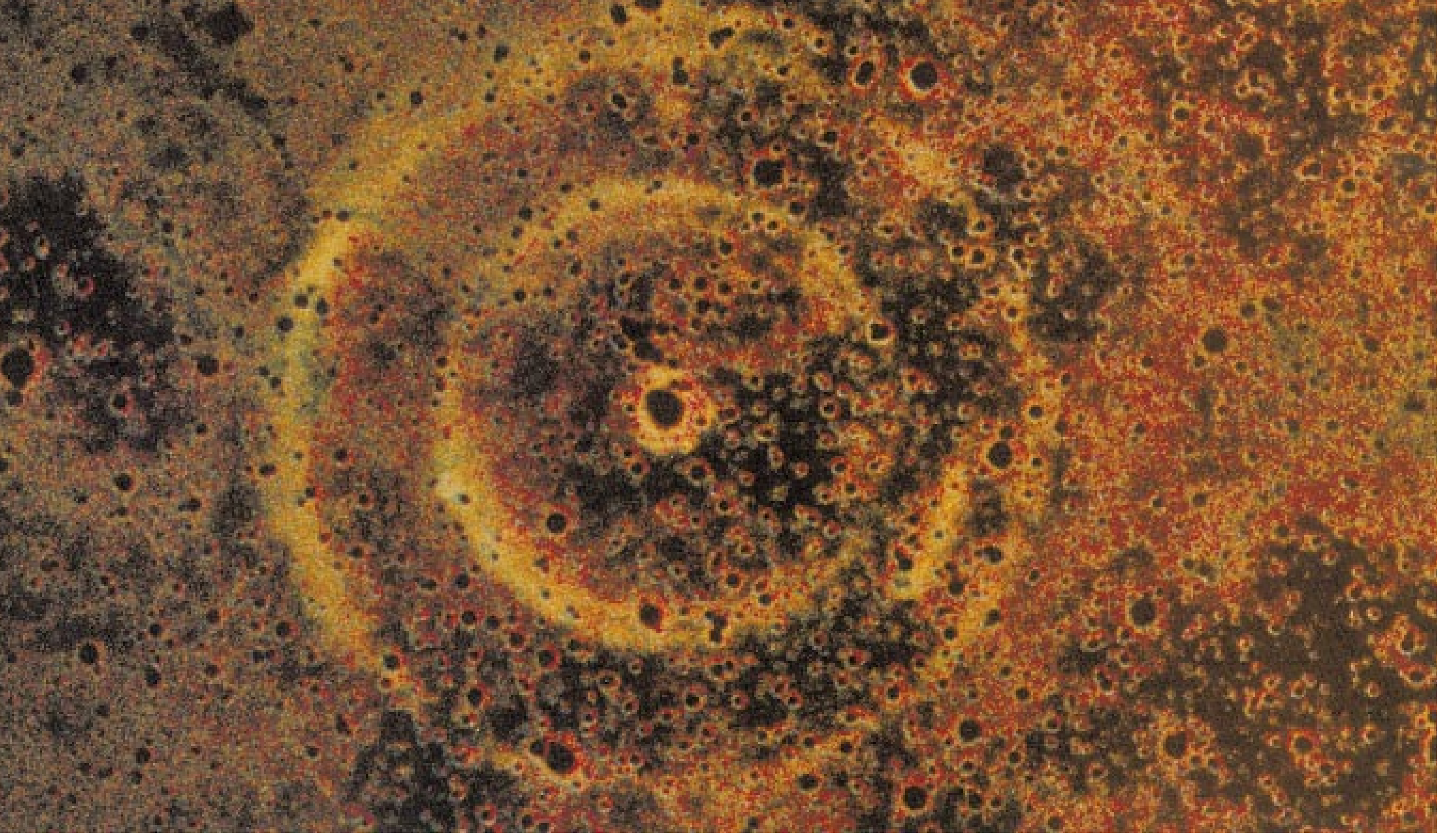
Color images present the information packed into an astronomical photograph in an attractive and intuitively obvious way. Indeed, measuring the colors of nebulae, stars and galaxies is central to understanding their composition and physical state. Yet researchers have been slow to appreciate the value of revealing the color inherent in many astronomical objects. Many astronomers consider color an abstract concept, easy to quantify but rarely rendered correctly on film.

The limitations of conventional color films have



DARK INTERSTELLAR CLOUDS (*top*), known as Bok globules, seem to be scattered throughout this nebular complex in the constellation Centaurus. The globules appear dark because they are silhouetted against a bright background. The Scorpius Cloud (*middle*) shows what happens to a dark nebula once hot, blue stars begin to form nearby. Radiation from a cluster of such stars, which lie just beyond the lower left edge of this photograph, is sweeping back and destroying the cloud. The stars' rays have given rise to a red zone of luminescent hydrogen around the Scorpius Cloud's edge. The dusty cloud itself contains a few embedded bright stars, visible by their reflected blue light. Eventually the nebula will disperse, leaving a loose gathering of young stars, somewhat like the open cluster NGC 3293 (*bottom*). One of the stars here has evolved into a bloated red giant. Such stars are red only in astronomical nomenclature; this image, in which the color balance is set like that for daylight color film, shows that red giants would actually appear yellow if the human eye were sufficiently sensitive to dim light.

DAVID F. MALIN is one of the leading practitioners of color astrophotography. He has published extensively on both photography and astronomy and has used innovative photographic techniques to discover exceptionally faint galaxies and nebulae. Malin trained as a chemist and spent many years in England using microscopes and x-rays to solve chemical problems. In 1975 he joined the Anglo-Australian Observatory in Epping, New South Wales, where he turned his attention to somewhat larger objects. He is now the observatory's photographic scientist. Malin's work has earned him numerous awards, including an honorary doctorate from the University of Sydney.



LUMINOUS RINGS (*above*) are echoes from supernova 1987A, a stellar explosion in one of the Milky Way's satellite galaxies. As the brief flash of light from the supernova raced outward, it passed through two sheets of dusty particles. The rings represent light that scattered off the dust; their tint reflects the color of the supernova at its brightest. This picture was made by subtracting a set of three-color images taken before the supernova appeared from another set taken after it vanished. Nebulosity common to both sets of plates did not cancel completely, producing the mottled background. The Crab nebula (*below*) consists of the expanding remains from a supernova explosion that became visible in the year 1054. This image was made by reprocessing photographic plates taken in 1955 and 1956 on the 200-inch Hale telescope. The almost textureless blue regions display radiation generated by high-speed charged particles moving through a magnetic field. The red filamentary structure, seen here at unprecedented clarity, consists of more conventional emission from hydrogen atoms.

consistently thwarted astronomers' interest in color photography. Modern films incorporate three photographic layers, which contain a blue-, green- or red-sensitive layer of silver halide; these layers in turn create cyan, yellow or magenta dyes when the film is processed. Such films are convenient to use, but they perform poorly at depicting the true colors of distant celestial objects.

Color films are also badly affected by the airglow of the night sky, a feeble but pervasive luminescence generated in the upper atmosphere by the recombination of ions and electrons split apart by sunlight during the day. Airglow causes a gradual fogging of photographic plates, limiting their sensitivity. Furthermore, some nebulae emit light only at certain discrete colors, or wavelengths, that are not well recorded by color films. Finally, color films do not respond reliably to the complex hypersensitization processes that astronomers routinely rely on to improve the sensitivity of the photographic medium.

To avoid these problems, my colleagues and I turned to Maxwell's venerable technique for creating color images. While investigating the nature of vision, Maxwell showed that he could replicate the color of a scene by photographing it three times, in red, green and blue light. From the resulting negatives, he made three positive, black-and-white slides, which he then projected onto a white screen through red, green or blue filters. By superimposing the projected images on top of one another, Maxwell demonstrated the world's first color photograph to an astonished audience at London's Royal Institution in 1861.

In the modernized version of Maxwell's process, we build a color photograph out of three separate black-and-white plates. Creating the image in this manner avoids many of the limitations of color film; it also permits the application of several powerful darkroom techniques at the stage when black-and-white negatives from the telescope are transformed into positive films.

Among the most valuable of these techniques is unsharp masking, which uses a blurred photographic positive to modulate the wide contrast range of some astronomical negatives without losing the fine detail. Another process, known as photographic amplification, selectively highlights the upper layers of the photographic emulsion, which tend to contain the faintest parts of the image. We can also stack together multiple negatives or even subtract images from one another (by illuminating a negative through a positive of the same object) in order to bring out the subtle variations among different exposures of the same patch of sky.

Once we are satisfied with the quality of the photographic positives, we place them in an enlarger and project them through the appropriate red, green or blue filter. Unlike Maxwell, we shine the images onto a large sheet of conventional color negative film, which becomes the master negative from which all subsequent color pictures are made.

That additive process enables us to control the color balance so that astronomical objects appear as they would to the human eye if it maintained its full color sensitivity at low light levels. Many of our targets have never before been seen in color, so their appearance often comes as a glorious surprise. The photographs on these pages illustrate both the scientific power and the aesthetic merit of our approach.

COLLIDING GALAXIES known as the Antennae (*right*) display two elegant curved streams of stars extruded by the galaxies' mutual gravitation. Blue regions denote the locations of vast clumps of hot, young stars whose formation was triggered by the encounter. The brighter, more intense colors in the upper galaxy suggest that it contained more gas than its companion when the two galaxies met. Dark dust clouds and pink hydrogen emission nebulae lie strewn throughout the disrupted region between the two galaxies. The Trifid nebula (*below*) is a much smaller example of such a nebula lying within our own galaxy. A cluster of infant stars resides where the three dark dust lanes appear to meet. Ultraviolet radiation from the hottest of these stars excites nearby hydrogen atoms, causing them to emit a ruddy light. Beyond the red region, some blue light penetrates and bounces off the dust particles that surround the nebula. The independent reflection nebula located to the right of the Trifid nebula has never before been captured in such detail in a color photograph.

Mastering Chaos

It is now possible to control some systems that behave chaotically. Engineers can use chaos to stabilize lasers, electronic circuits and even the hearts of animals

by William L. Ditto and Louis M. Pecora

What good is chaos? Some would say it is unreliable, uncontrollable and therefore unusable. Indeed, no one can ever predict exactly how a chaotic system will behave over long periods. For that reason, engineers have typically dealt with chaos in just one way: they have avoided it. We find that strategy somewhat shortsighted. Within the past few years we and our colleagues have demonstrated that chaos is manageable, exploitable and even invaluable.

Chaos has already been applied to increase the power of lasers, synchronize the output of electronic circuits, control oscillations in chemical reactions, stabilize the erratic beat of unhealthy animal hearts and encode electronic messages for secure communications. We anticipate that in the near future engineers will no longer shun chaos but will embrace it.

There are at least two reasons why chaos is so useful. First, the behavior of a chaotic system is a collection of many orderly behaviors, none of which dominates under ordinary circumstances. In recent years, investigators have shown that by perturbing a chaotic system in the right way, they can encourage the system to follow one of its many regular behaviors. Chaotic systems are unusually flexible because they can rapidly switch among many different behaviors.

Second, although chaos is unpredictable, it is deterministic. If two nearly identical chaotic systems of the appropriate type are impelled, or driven, by the same signal, they will produce the same output, even though no one can say what that output might be [see "The Amateur Scientist," page 120]. This phenomenon has already made possible a variety of interesting technologies for communications.

For more than a century, chaos has been studied almost exclusively by a few theoreticians, and they must be credited for developing some of the concepts on which all applications are based. Most natural systems are nonlinear: a change in behavior is not a simple function of a change in conditions. Chaos is one type of nonlinear behavior. The distinguishing feature of chaotic systems is that they exhibit a sensitivity to initial conditions. To be more specific, if two chaotic systems that are nearly identical are in two slightly different states, they will rapidly evolve toward very different states.

To the casual observer, chaotic systems appear to behave in a random fashion. Yet close examination shows that they have an underlying order. To visualize dynamics in any system, the Irish-born physicist William Hamilton and the German mathematician Karl Jacobi and their contemporaries devised, more than 150 years ago, one of the fundamental concepts necessary for understanding nonlinear dynamics: the notion of state space.

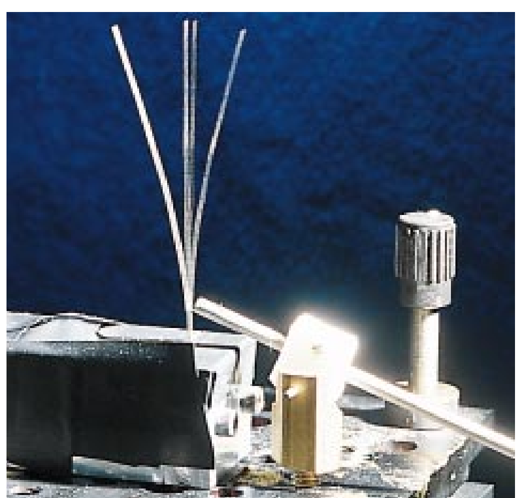
Any chaotic system that can be described by a mathematical equation includes two kinds of variables: dynamic and static. Dynamic variables are the fundamental quantities that are changing all the time. For a chaotic mechanism, the dynamic variables might be the position of a moving part and its velocity. Static variables, which might also be called parameters, are set at some point but then are never changed. The static variable of a chaotic mechanism might be the length of some part or the speed of a motor.

State space is essentially a graph in which each axis is associated with one dynamic variable. A point in state space represents the state of the system at a given time. As the system changes, it moves from point to point in state space, defining a trajectory, or curve. This trajectory represents the history of the dynamic system.

Chaotic systems have complicated trajectories in state space. In contrast, linear systems have simple trajectories, such as loops [*see box on page 80*]. Yet

WILLIAM L. DITTO and LOUIS M. PECORA have pioneered methods for controlling the chaotic behavior of mechanical, electrical and biological systems. Ditto is assistant professor of physics at the Georgia Institute of Technology. He won the Office of Naval Research Young Investigator Award. Ditto and his collaborators were the first to control chaos in an experimental system. Since 1977 Pecora has been a research physicist at the U.S. Naval Research Laboratory and now heads a program in nonlinear dynamics in solid-state systems there. Pecora is in the process of patenting four devices that exploit chaos.

METALLIC RIBBON sways chaotically in a magnetic field whose strength fluctuates periodically. Recent advances make it possible to control such chaotic motions. The ribbon is made of a material whose stiffness depends on the strength of the field. The magnetic field changes at a rate of around one cycle per second. A strobe light flashes at the same frequency so that long-exposure photography captures the irregularities in the motion of the ribbon. A computer system analyzes the movement of the ribbon using a scheme developed by theorists at the University of Maryland. By making slight changes to the field, the computer system can transform the chaotic motions (*right*) to periodic oscillations of almost any frequency (*below*).

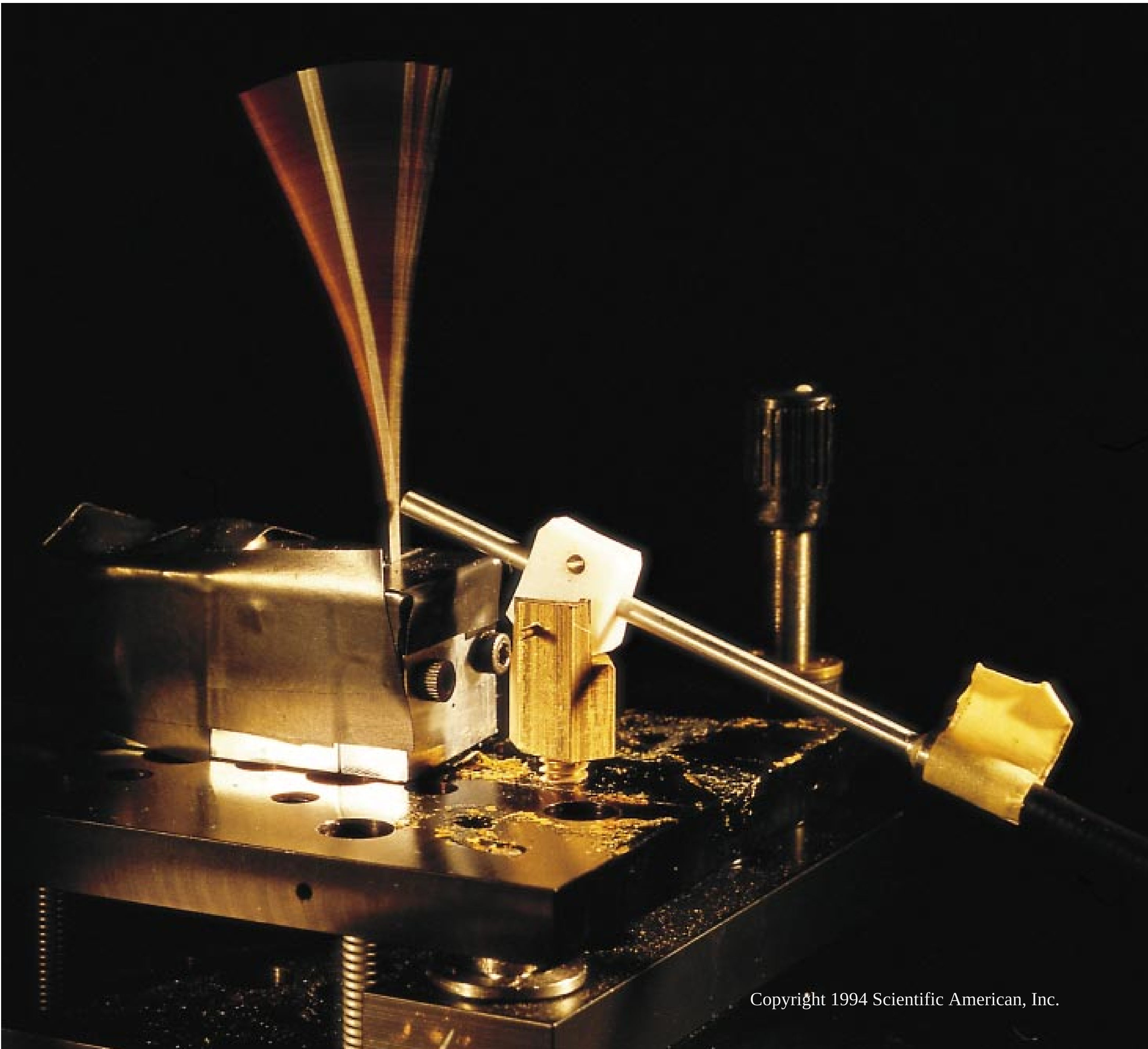


the trajectory of a chaotic system is not random; it passes through certain regions of state space while avoiding others. The trajectory is drawn toward a so-called chaotic attractor, which in some sense is the very essence of a chaotic system. The chaotic attractor is the manifestation of the fixed parameters and equations that determine the values of the dynamic variables.

So if one measures the trajectory of a chaotic system, one cannot predict where it will be on the attractor at some point in the distant future. The chaotic attractor, on the other hand, remains the same no matter when one measures it. Once researchers have obtained information about the chaotic attractor of a system, they can begin to use chaos to their advantage.

In 1989 one of us (Pecora) discovered that a chaotic system could be built in such a way that its parts would act in perfect synchrony. As is well known, isolated chaotic systems cannot synchronize. If one could construct two chaotic systems that were virtually identical but separate, they would quickly fall out of step with each other because any slight difference between the two systems would be magnified. In some cases, however, the parts of a system can be arranged so that they exhibit identical chaotic behavior. In fact, these parts can be located at great distances from one another, making it possible to use synchronized chaos for communications.

To build a chaotic system whose parts are synchronous, one needs to understand the concept of stability. A system is stable if, when perturbed somewhat, its trajectory through state space changes only a little from what it might have been otherwise. The Russian mathematician Aleksandr M. Lyapunov realized that a single number could be used to represent the change caused by a perturbation. He divided the size of the perturbation at one instant in time by its size a moment before. He then performed the same computation at various intervals and averaged the results.



This quantity, now known as the Lyapunov multiplier, describes how much, on average, a perturbation will change. If the Lyapunov multiplier is less than one, the perturbations die out, and the system is stable. If the multiplier is greater than one, however, the disturbances grow, and the system is unstable. All chaotic systems have a Lyapunov multiplier greater than one and so are always unstable. This is the root of the unpredictability of chaos.

The parts of a chaotic system must be stable if they are to be synchronized. This does not mean that the entire system cannot be chaotic. If two similar, stable parts are driven by the same chaotic signal, they will both seem to exhibit chaotic behavior, but they will suppress rather than magnify any differences between them, thereby creating an opportunity for synchronized chaos.

One design developed by Pecora and Thomas L. Carroll of the U.S. Naval Research Laboratory uses a subsystem of an ordinary chaotic system as the synchronizing mechanism. They first distinguished the synchronizing subsystem from other supporting parts and then duplicated the synchronizing subsystems [*see illustrations on opposite page*]. The supporting subsystem then supplied a driving signal for the original synchronizing subsystem and its duplicate.

If the synchronizing subsystems have Lyapunov multipliers that are less than one, that is, if these subsystems are stable, they will behave chaotically but will be in complete synchrony. The stability of the subsystems guarantees that any small perturbations will be damped out, and therefore the synchronizing subsystems will react to the signals from the supporting subsystem in practically the same way, no matter how complex the signals are.

To demonstrate synchronized chaos initially, Pecora created a computer simulation based on the chaotic Lorenz system, which is named after American meteorologist Edward N. Lorenz, who in 1963 discovered chaotic behavior in a computer study of the weather. The Lorenz system has three dynamic variables, and consequently the state-space picture of such systems is three-dimensional. Plotting the trajectory of the Lorenz system in state space reveals what is known as the Lorenz chaotic attractor [see "Chaos," by James P. Crutchfield, J. Doyne Farmer, Norman H. Packard and Robert S. Shaw; SCIENTIFIC AMERICAN, December 1986].

For the computer simulation, Pecora started with the three dynamic variables of the Lorenz system and chose one of them as the driving signal. The subsystem consisting of the two remaining dynamic variables was then duplicated. Although the subsystem and its duplicate were initially set up to produce different outputs, they quickly converged, generating chaotic signals in synchrony. In state space the two subsystems began at different points, caught up to each other and waltzed about on their chaotic attractors in a synchronized ballet.

Until this discovery, scientists had no reason to believe that the stability of a subsystem could be independent of the stability of the rest of the system. Nor

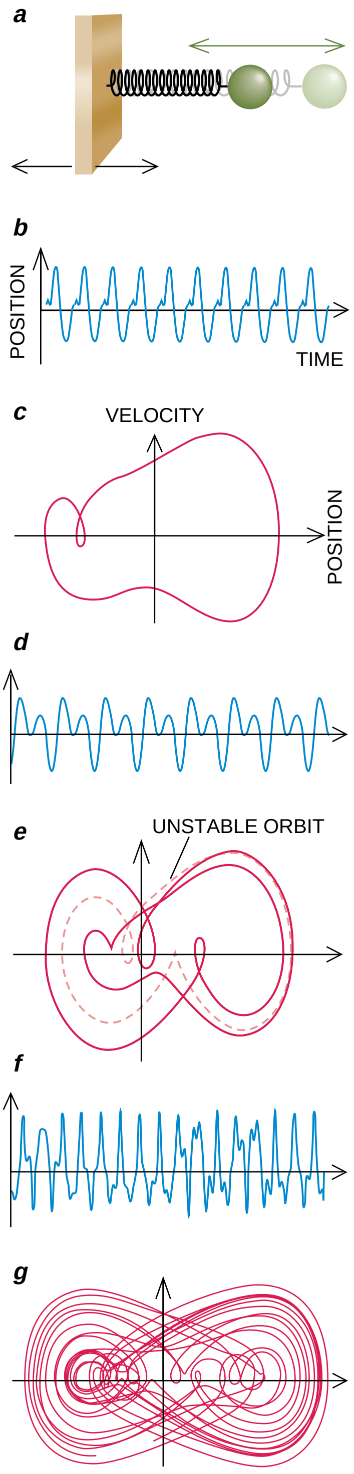
Representing Chaos

Conventionally, the motion of an object is represented as a "time series," a graph showing the change in position over time. Chaotic motions, however, can often be more conveniently visualized in "state space," a plot of the history of the changing variables, which are typically the position and velocity of the object. Consider, for example, the motion of the ball in the mechanism shown at the right (*a*). The ball is attached to a spring that gets stiffer as it is stretched or compressed. (In technical terms, the spring delivers a nonlinear restoring force.) As the board moves cyclically back and forth, the spring is pushed and pulled and causes the ball to move. Hence, the movement of the ball ultimately depends on the force with which the spring is pushed.

If the force is weak, the ball moves in a simple trajectory, which repeats with each cycle of the force from the board (*b*). The state-space path (*c*) shows all the information about the motion of the ball at each instant of time. Because the motion is periodic, the path will retrace itself with each board cycle. The state space reveals a so-called period-one attractor. The same kind of behavior would be observed if the ball were attached to an ordinary (linear) spring.

The interesting behavior occurs as the driving force on the spring is increased. At a certain point, the ball can be made to move back and forth in a more complicated way for each oscillation of the board. Hence, the time series and state-space path change (*d, e*). Specifically, the previous period-one orbit becomes unstable, and the system produces a period-two attractor, that is, it takes two board cycles before the path retraces itself.

If the force on the spring increases past some threshold, the ball moves chaotically. No pattern is apparent in the time series (*f*). The state space, on the other hand, reveals a chaotic attractor (*g*): the state-space path never retraces itself, yet it occupies only certain regions of state space. Indeed, a chaotic attractor can be considered a combination of unstable periodic orbits.



had anyone thought that a nonlinear system could be stable when driven with a chaotic signal.

Stability depends not only on the properties of the subsystem itself but on the driving signal. A subsystem may be stable when driven by one type of chaotic signal but not when driven by another. The trick is to find those subsystems that react to a chaotic signal in a stable way. In some cases, the stability of a subsystem can be estimated using a mathematical model, but in general such predictions are difficult.

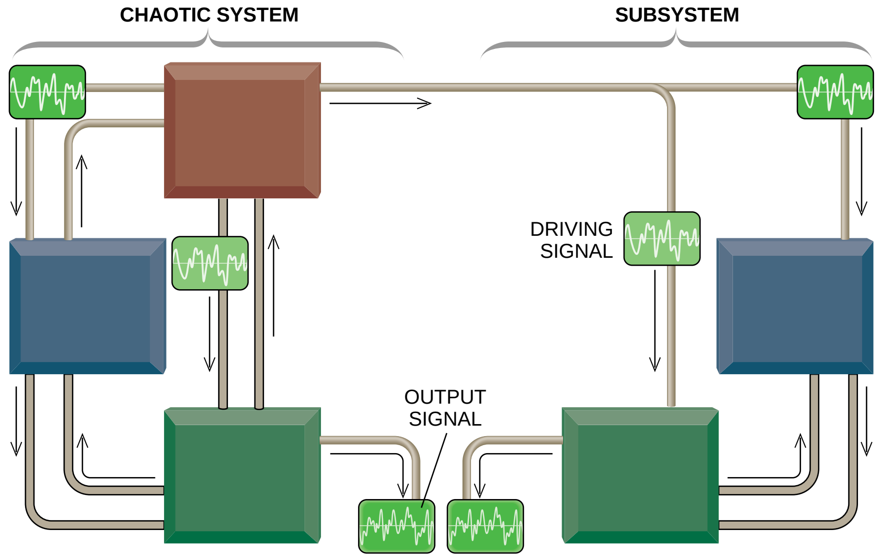
Investigators have only begun to experiment with synchronized chaos. In 1989 Carroll built the first synchronized chaos circuit. By duplicating part of a circuit that exhibited chaotic behavior, he created a circuit in which the subsystem and its duplicate were driven by the same chaotic signal. These two subsystems generated voltages that fluctuated in a truly chaotic fashion but were always in step with each other.

Carroll realized that his synchronized chaos circuit might be used for a private communications system. For example, imagine that Bill wants to send a secret message to Al. Bill has a device that generates a chaotic drive signal, and he has a subsystem that reacts to the signal in a stable way. Al has a copy of the subsystem. Bill translates his secret message into an electronic signal and combines it with the chaotic output of his subsystem. Bill then transmits the encoded message as well as the drive signal.

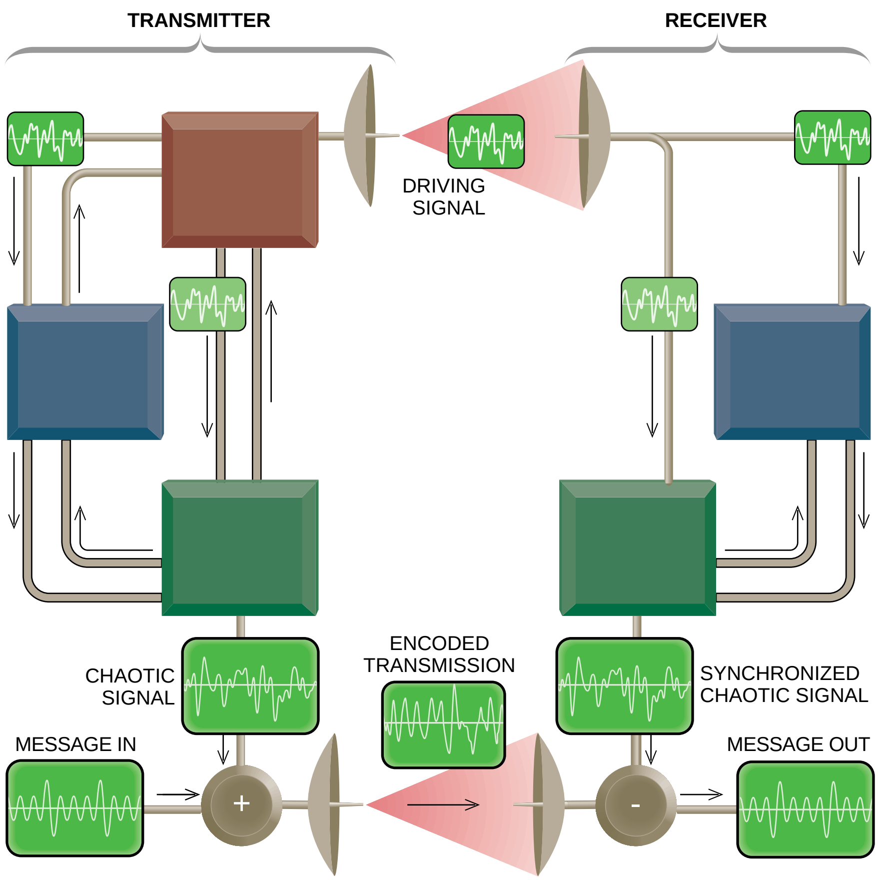
Anyone who intercepts these signals will detect only chaotic noise and will be unable to extract any information (unless, of course, he or she manages to get a copy of Bill's or Al's subsystem). When Al receives the drive signal, he sends it through his subsystem, which reproduces the chaotic output of Bill's subsystem. Al can then subtract this output from the encoded information signal to recover the secret message.

It is rather easy, as Carroll has demonstrated, to recover signals buried in chaos. To build a practical, secure communications system, however, engineers will probably need to develop sophisticated schemes in which the encoding process involves more than just adding chaos to message signals.

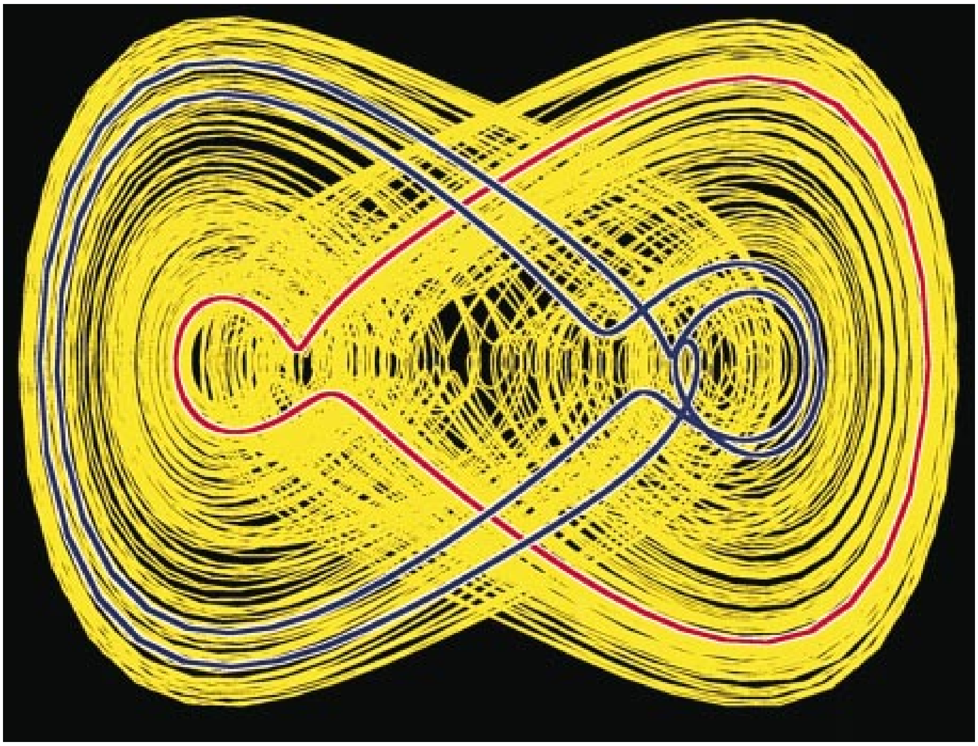
Researchers at the Massachusetts Institute of Technology, Washington State University and the University of California at Berkeley are building new combinations of chaotic subsystems for signal processing and communications. We have also continued our work and found a chaotic system that performs the same operations as a phase-locked



CHAOTIC SIGNALS produced by a system and a subsystem can be generated in complete synchrony. The important feature of the system is that the blue and green components are stable when driven by the chaotic signals from the red part. In other words, small changes in the initial settings of the blue and green parts have little or no effect on the behavior they eventually settle into. The subsystem consists of duplicates of the blue and green components of the system. Although the red, blue and green parts of the system send signals to one another, there is no feedback between the red component and the subsystem.



PRIVATE COMMUNICATIONS can be achieved using the scheme for synchronized chaos [*see upper illustration on this page*]. The transmitter adds a chaotic signal to a message signal; it then broadcasts the encoded signal and a drive signal. The drive signal is sent to two components in the receiver, causing them to generate a chaotic signal in synchrony with the signal produced in the transmitter. The chaotic signal is then subtracted from the encoded signal to yield the original information.



CHAOTIC ATTRACTOR consists of many periodic orbits—for example, the period-one orbit (*red*) and the period-two orbit (*blue*). This attractor represents a system whose velocity and position change along a single direction. One axis represents position, and the other is velocity. Attractors may be multidimensional because systems can have many different state-space variables, that is, positions and velocities that vary in three dimensions.

loop—a device that, among other things, allows an FM radio receiver to track the changes in the transmitted signal.

Carroll and Pecora had such success with using chaotic signals to drive stable parts of chaotic systems that they suspected there might be an advantage to employing chaotic signals to drive stable periodic systems. Their intuition led them to another series of promising applications. Imagine two systems that are driven by the same periodic signal and have the same nonchaotic period-two attractor. The term "period two" means that in two periods of the driving signal, the system travels around the attractor once. If two such systems are driven with the same signal but are initially at opposite ends of the attractor, they will cycle around the attractor and never catch up to each other. In other words, if the two systems start out of phase, they will remain out of phase forever.

By changing the periodic driving signal to a certain type of chaotic signal, workers have recently discovered that two systems can be coaxed to operate in phase. Yet not all types of chaotic signals will perform this task. Some such signals may cause the system themselves to behave chaotically, thereby eliminating any chance of getting them in phase. Engineers must therefore figure out what type of chaotic signal to use for each kind of system.

A good first choice is the signal produced by a system originally invented by Otto E. Rössler. The Rössler signal is nearly periodic and resembles a sine wave whose amplitude and wavelength have been randomized somewhat from one cycle to the next. Such signals, which have come to be known as pseudoperiodic, can be made, in general, simply by taking a periodic signal and adding chaos to it. If several identical systems whose attractors are all period two or greater are driven by the appropriate pseudoperiodic signal, they will all get in step.

This application was first demonstrated by Carroll in 1990. He built a set of electronic circuits that had period-two attractors and drove them with one of several pseudoperiodic signals, including the Rössler type. In each case, he found he was able to solve the out-of-step problem. Yet the circuits did get out of phase for short periods. That occurred because it is impossible to build several circuits that are exactly the same. Nevertheless, it is quite feasible to keep the circuits in phase 90 percent of the time.

Whereas some investigators have searched for uses of chaos, others have sought to control it. One key to this effort lies in the realization that a chaotic attractor is an infinite collection of unstable periodic behaviors. The easiest way to illustrate this point is to consider a system consisting of a weight, a "nonlinear" spring and a motor. One end of the spring hangs from the motor; the other is attached to the weight. The motor lifts the spring up and down with a force that can be adjusted. The dynamic variables of the system are the position of the weight and its velocity, and these define the state space. If the driving force of the motor is weak, the weight will move up and down once for each cycle of the motor. At the same time, the weight speeds up and slows down, and therefore the trajectory in state space is a single loop, or what is technically called a period-one orbit.

If the driving force of the motor is increased somewhat, the period-one orbit becomes an unstable behavior, and the weight will move up and down once for every two cycles of the motor. In state space the trajectory is now a double loop, or period-two orbit. If the driving force is increased again, the period-two orbit becomes unstable, and a period-four orbit will emerge. Indeed, if the driving force is sufficiently strong, all periodic orbits become unstable, and a chaotic attractor will appear. In a rigorous sense, the chaotic attractor is an ensemble of unstable periodic orbits.

Four years ago Edward Ott, Celso Grebogi and James A. Yorke of the University of Maryland developed a scheme in which a chaotic system can be encouraged to follow one particular unstable orbit through state space. The scheme can therefore take advantage of the vast array of possible behaviors that make up the chaotic system.

The method devised by Ott, Grebogi and Yorke (referred to as OGY) is conceptually straightforward. To start, one obtains information about the chaotic system by analyzing a slice of the chaotic attractor. After the information about this so-called Poincaré section has been gathered, one allows the system to run and waits until it comes near a desired periodic orbit in the section. Next the system is encouraged to remain on that orbit by perturbing the appropriate parameter. One strength of this method is that it does not require a detailed model of the chaotic system but only some information about the Poincaré section. It is for this reason that the method has been so successful in controlling a wide variety of chaotic systems.

POINCARÉ SECTION is, more or less, a perpendicular slice through a chaotic attractor, in this case, a three-dimensional version of the attractor shown on the opposite page. The points of the Poincaré section represent different unstable periodic orbits. The period-one orbit is a single point (*red*), and the period-two orbit is two points (*blue*). The determination of the Poincaré section is a key step in controlling chaos.

The practical difficulties of the OGY method involve getting the Poincaré section and then calculating the control perturbations. One of the simplest ways to obtain a Poincaré section is to measure, at some regular interval, the position of the state-space trajectory of the system. For example, the measurements could be made at the end of every cycle of the driving signal. This technique produces a map in which a period-one orbit appears as one point, a period-two orbit appears as two points and so on.

To calculate the control perturbations, one analyzes the Poincaré section to determine how the system approaches the desired orbit, or fixed point. This analysis requires three steps. First, one determines the directions along which the system converges toward or diverges away from the fixed point. Second, one must obtain the rate of convergence or divergence—a quantity that is related to the Lyapunov multiplier. Last, one must figure out how much the orbit shifts in phase space if the control parameter is changed in one way or another. This three-step calculation provides the information necessary to determine the perturbations that will nudge the system toward the desired periodic orbit.

When the control parameter is actually changed, the chaotic attractor is shifted and distorted somewhat. If all goes according to plan, the new attractor encourages the system to continue on the desired trajectory. If the system starts to go astray, the control parameter is changed again, producing yet another attractor with the desired properties.

The process is similar to balancing a marble on a saddle. If the marble is initially placed in the center of the saddle, it tends to roll off one side or the other but is unlikely to fall off the front or back. To keep the marble from rolling off, one needs to move the saddle quickly from side to side. Likewise, one needs to shift the attractor to compensate for the system's tendency to fly off the desired trajectory in one direction or another. And just as the marble reacts to small movements of the saddle, the trajectory is very sensitive to changes in the attractor.

What is truly extraordinary about the work of Ott, Grebogi and Yorke is that it showed that the presence of chaos could be an advantage in controlling dynamic behavior. Because chaotic systems are extremely sensitive to initial conditions, they also react very rapidly to implemented controls.

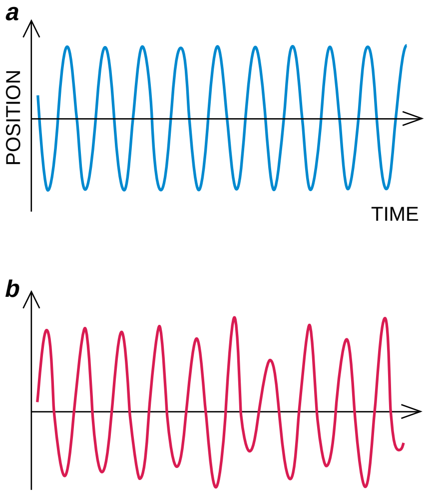
In 1990 one of us (Ditto), along with Mark L. Spano and Steven N. Rauseo of the Naval Surface Warfare Center, set out to test the ideas of Ott, Grebogi and Yorke. We did not expect things to work right away. Few theories ever survive initial contact with experiment. But the work of Ott, Grebogi and Yorke proved to be an exception. Within a couple of months we achieved control over chaos in a rather simple experimental system.

The experiment requires a metallic ribbon whose stiffness can be changed by applying a magnetic field. The bottom end of the ribbon is clamped to a base; the top flops over either to the left or right. When the ribbon is exposed to a field whose strength is varied periodically at a rate around one cycle per second, the ribbon buckles chaotically. A second magnetic field served as the control parameter.

Our goal was to change the chaotic motion of the ribbon to a periodic one by applying the OGY method. We first needed to obtain the Poincaré map of the system's chaotic attractor. The map was created by recording the position of the ribbon once per drive cycle, that is, the frequency with which the magnetic field was modulated. A computer stored and analyzed the map; it then calculated how the control parameter should be altered so that the system would follow a period-one orbit through state space. We found that even in the presence of noise and imprecise measurements, the ribbon could easily be controlled about an unstable period-one orbit in the chaotic region. We were genuinely surprised at how easy it was to implement and exploit the chaos.

The ribbon remained under control without a failure for three days, after which we got bored and wanted to attempt something else. We tried to make the system fail by adding external noise, altering the parameters and demonstrating the system to visitors (a sure way to make any experiment fail). In addition, we succeeded in forcing the system to follow period-one, period-two and period-four orbits, and we could switch between the behaviors at will.

As often happens in science, discoveries lead in unexpected directions. Earle R. Hunt of Ohio University took notice of our experiments and decided to attempt to control chaos in an electronic circuit. Hunt's circuit was made out of readily available components. He was able to coax the system into orbits with periods as high as 23 and at drive frequencies as high as 50,000 cycles per second. Not only did he prove that chaos could be controlled in electronic circuits, but he also showed that



NONCHAOTIC SYSTEMS that are out of phase can sometimes be made to operate in phase if they are driven with chaotic signals. For example, if two period-two systems are driven with the same periodic signal (*a*) and start out of phase, they will remain out of phase. If the same two systems are driven with a Rössler signal (*b*), which is mildly chaotic, they can be coaxed to operate in phase.

scientific advances can be made on a low budget.

Hunt's chaos controller—which employs a variation of the OGY method—has the advantage that it eliminates the need to obtain a Poincaré section; information for the controller can be obtained directly from measurements of the chaotic behavior. Hunt's circuit enables the manipulation of chaos in rapidly changing systems. Armed with these results, Hunt attended, in October 1991, the First Experimental Chaos Conference, having solved a problem he did not know existed.

Hunt had unknowingly overcome a technological obstacle that had stymied Rajarshi Roy and his colleagues at the Georgia Institute of Technology. Roy had been studying the effects of a "doubling crystal" on laser light. The crystal doubles the frequency of the incident light or, equivalently, halves the wavelength. The laser used by Roy generated infrared light at a wavelength of 1,064 nanometers, and the crystal converted the light to green at a wavelength of 532 nanometers. The crystal did not work perfectly, however. If the doubling crystal was oriented in certain ways, the intensity of green light would fluctuate chaotically.

Roy was eager to control the chaos in the laser system, but he knew the OGY method would not work, because it was difficult and impractical to obtain the Poincaré section for the laser system. Furthermore, Roy was skeptical that a computer system could perform calculations fast enough to implement the OGY method. Fortunately, he had learned about Hunt's work at the conference. Less than two weeks later Roy constructed a prototype controller for their laser, and to their astonishment, it worked on the first try. Within days they were controlling periods as high as 23 and at driving frequencies of 150,000 cycles per second.

Roy demonstrated that he could control chaotic fluctuations in the laser intensity and could stabilize unstable high-period oscillations. The energy of the laser could therefore be dumped into desired frequencies instead of being distributed across a broad band of frequencies. (Roy has also employed this technique to eliminate almost completely intensity fluctuations in his laser system.) Consequently, investigators can design laser systems with more flexibility and stability than ever before.

The field of laser research has also benefited from other advances in chaos control. For example, the team of Ira B. Schwartz, Ioana A. Triandaf, Carroll and Pecora at the Naval Research Laboratory developed a method known as tracking to extend the range over which the control of chaos can be maintained. Tracking compensates for parameters that change as the system ages or that slowly drift for one reason or another.

In recent months, tracking has been applied to both chaotic circuits and lasers with astounding results. For example, the laser remains stable only for a very limited power range if the parameters of the control mechanism do not adapt. By using tracking, researchers can maintain control over a much wider power range, and amazingly, they find they can increase the output power by a factor of 15.

One problem with the OGY method is that an unacceptable amount of time may elapse as one waits for the system naturally to approach the desired orbit in the chaotic attractor. Troy Shinbrot and his colleagues at the University of Maryland and the Naval Surface Warfare Center have demonstrated a technique that rapidly moves the chaotic system from an arbitrary initial state to a desired orbit in the attractor. When Shinbrot and his co-workers implemented their method in the magnetic-ribbon experiment, they were able to reduce the time needed to acquire unstable orbits by factors as high as 25.

Another collaboration that emerged from the First Experimental Chaos Conference led to the first technique for controlling chaos in a biological system. The team—which included Ditto, Spano, Alan Garfinkel and James N. Weiss of the School of Medicine, University of California at Los Angeles—studied an isolated part of a rabbit heart. We were able to induce fast, irregular contractions of the heart muscle by injecting a drug called ouabain into the coronary arteries. Once this arrhythmia started, we stimulated the heart with electric signals that were generated according to a scheme we adapted from the OGY method. These seemingly random signals were sufficient to establish a regular beat and sometimes reduced the heart rate to normal levels. On the other hand, signals that were random or periodic did not stop the arrhythmia and often made it worse.

Investigators have already begun to test whether variations of the OGY method could be used to control arrhythmias in human hearts. They suspect the technique might work for atrial or ventricular fibrillation, in which upper or lower chambers of the heart contract irregularly and cease to pump blood effectively. In the near future, it may be possible to develop pacemakers and defibrillators that take advantage of techniques for controlling chaos.

Scientists and engineers have just begun to appreciate the advantages of designing devices to exploit, rather than avoid, nonlinearity and chaos. Whereas linear systems tend to do only one thing well, nonlinear devices may be capable of handling several tasks. Nonlinear applications promise more flexibility, faster response and unusual behaviors. As we continue to investigate the nonlinearity inherent in natural and physical systems, we may learn not just to live with chaos, not just to understand it, but to master it.

FURTHER READING

SYNCHRONIZATION IN CHAOTIC SYSTEMS. L. M. Pecora and T. L. Carroll in *Physical Review Letters,* Vol. 64, No. 8, pages 821–824; February 19, 1990.

CONTROLLING CHAOS. Edward Ott, Celso Grebogi and James A. Yorke in *Physical Review Letters,* Vol. 64, No. 11, pages 1196–1199; March 12, 1990.

EXPERIMENTAL CONTROL OF CHAOS. W. L. Ditto, S. N. Rauseo and M. L. Spano in *Physical Review Letters,* Vol. 65, No. 26, pages 3211–3214; December 24, 1990.

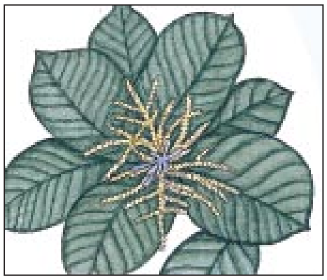
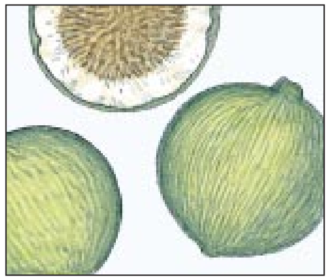
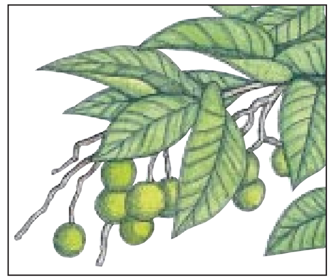
CONTROLLING CARDIAC CHAOS. A. Garfinkel, M. L. Spano, W. L. Ditto and J. N. Weiss in *Science,* Vol. 257, pages 1230–1235; August 28, 1992.

PROCEEDINGS OF THE 1ST EXPERIMENTAL CHAOS CONFERENCE. Edited by Sandeep Vohra, Mark Spano, Michael Shlesinger, Lou Pecora and William Ditto. World Scientific, 1992.

Diet and Primate Evolution

Many characteristics of modern primates, including our own species, derive from an early ancestor's practice of taking most of its food from the tropical canopy

by Katharine Milton

As recently as 20 years ago, the canopy of the tropical forest was regarded as an easy place for apes, monkeys and prosimians to find food. Extending an arm, it seemed, was virtually all our primate relatives had to do to acquire a ready supply of edibles in the form of leaves, flowers, fruits and other components of trees and vines. Since then, efforts to understand the reality of life for tree dwellers have helped overturn that misconception.

My own field studies have provided considerable evidence that obtaining adequate nutrition in the canopy—where primates evolved—is, in fact, quite difficult. This research, combined with complementary work by others, has led to another realization as well: the strategies early primates adopted to cope with the dietary challenges of the arboreal environment profoundly influenced the evolutionary trajectory of the primate order, particularly that of the anthropoids (monkeys, apes and humans).

Follow-up investigations indicate as well that foods eaten by humans today, especially those consumed in industrially advanced nations, bear little resemblance to the plant-based diets anthropoids have favored since their emergence. Such findings lend support to the suspicion that many health problems common in technologically advanced nations may result, at least in part, from a mismatch between the diets we now eat and those to which our bodies became adapted over millions of years. Overall, I would say that the collected evidence justifiably casts the evolutionary history of primates in largely dietary terms.

The story begins more than 55 million years ago, after angiosperm forests spread across the earth during the late Cretaceous (94 to 64 million years ago). At that time, some small, insect-eating mammal, which may have resembled a tree shrew, climbed into the trees, presumably in search of pollen-distributing insects. But its descendants came to rely substantially on edible plant parts from the canopy, a change that set the stage for the emergence of the primate order.

Natural selection strongly favors traits that enhance the efficiency of foraging. Hence, as plant foods assumed increasing importance over evolutionary time (thousands, indeed millions, of years), selection gradually gave rise to the suite of traits now regarded as characteristic of primates. Most of these traits facilitate movement and foraging in trees. For instance, selection yielded hands well suited for grasping slender branches and manipulating found delicacies.

Selective pressures also favored considerable enhancement of the visual apparatus (including depth perception, sharpened acuity and color vision), thereby helping primates travel rapidly through the three-dimensional space of the forest canopy and easily discern the presence of ripe fruits or tiny, young leaves. And such pressures favored increased behavioral flexibility as well as the ability to learn and remember the identity and locations of edible plant parts. Foraging benefits conferred by the enhancement of visual and cognitive skills, in turn, promoted development of an unusually large brain, a characteristic of primates since their inception.

As time passed, primates diverged into various lineages: first prosimians, most of which later went extinct, and then monkeys and apes. Each lineage arose initially in response to the pressures of a somewhat different dietary niche; distinct skills are required to become an efficient forager on a particular subset of foods in the forest canopy. Then new dietary pressures placed on some precursor of humans paved the way for the development of modern humans. To a great extent, then, we are truly what we eat.

My interest in the role of diet in primate evolution grew out of research I began in 1974. While trying to decide on a topic for my doctoral dissertation in physical anthropology, I visited the tropical forest on Barro Colorado Island in the Republic of Panama. Studies done on mantled howler monkeys (*Alouatta palliata*) in

KATHARINE MILTON is professor of anthropology at the University of California, Berkeley. After earning her doctorate from New York University in 1977, she spent three years conducting field studies on the foraging behavior of howler and spider monkeys. She joined the Berkeley faculty in 1980. Milton is now studying the dietary behavior and diseases of indigenous human populations living in the tropical forests of Brazil.

the 1930s at that very locale had inadvertently helped foster the impression that primates enjoyed the "life of Riley" in the canopy.

Yet, during my early weeks of following howlers, I realized they were not behaving as expected. Instead of sitting in a tree and eating whatever happened to be growing nearby, they went out of their way to seek specific foods, meanwhile rejecting any number of seemingly promising candidates. Having found a preferred food, they did not sate themselves. Instead they seemed driven to obtain a mixture of leaves and fruits, drawn from many plant species.

The old easy-living dogma was clearly far too simplistic. I decided on the spot to learn more about the problems howlers and other anthropoids face meeting their nutritional needs in the tropical forest. I hoped, too, to discern some of the strategies they had evolved to cope with these dietary difficulties.

The challenges take many forms. Because plants cannot run from hungry predators, they have developed other defenses to avoid the loss of their edible components. These protections include a vast array of chemicals known as secondary compounds (such as tannins, alkaloids and terpenoids). At best, these chemicals taste awful; at worst, they are lethal.

Also, plant cells are encased by walls made up of materials collectively referred to as fiber or roughage: substances that resist breakdown by mammalian digestive enzymes. Among the fibrous constituents of the cell wall are the structural carbohydrates—cellulose and hemicellulose—and a substance called lignin; together these materials give plant cell walls their shape, hardness and strength. Excessive intake of fiber is troublesome, because when fiber goes undigested, it provides no energy for the feeder. It also takes up space in the gut. Hence, until it can be excreted, it prevents intake of more nourishing items. As will be seen, many primates, including humans, manage to extract a certain amount of energy, or calories, from fiber despite their lack of fiber-degrading enzymes. But the process is time-consuming and thus potentially problematic.

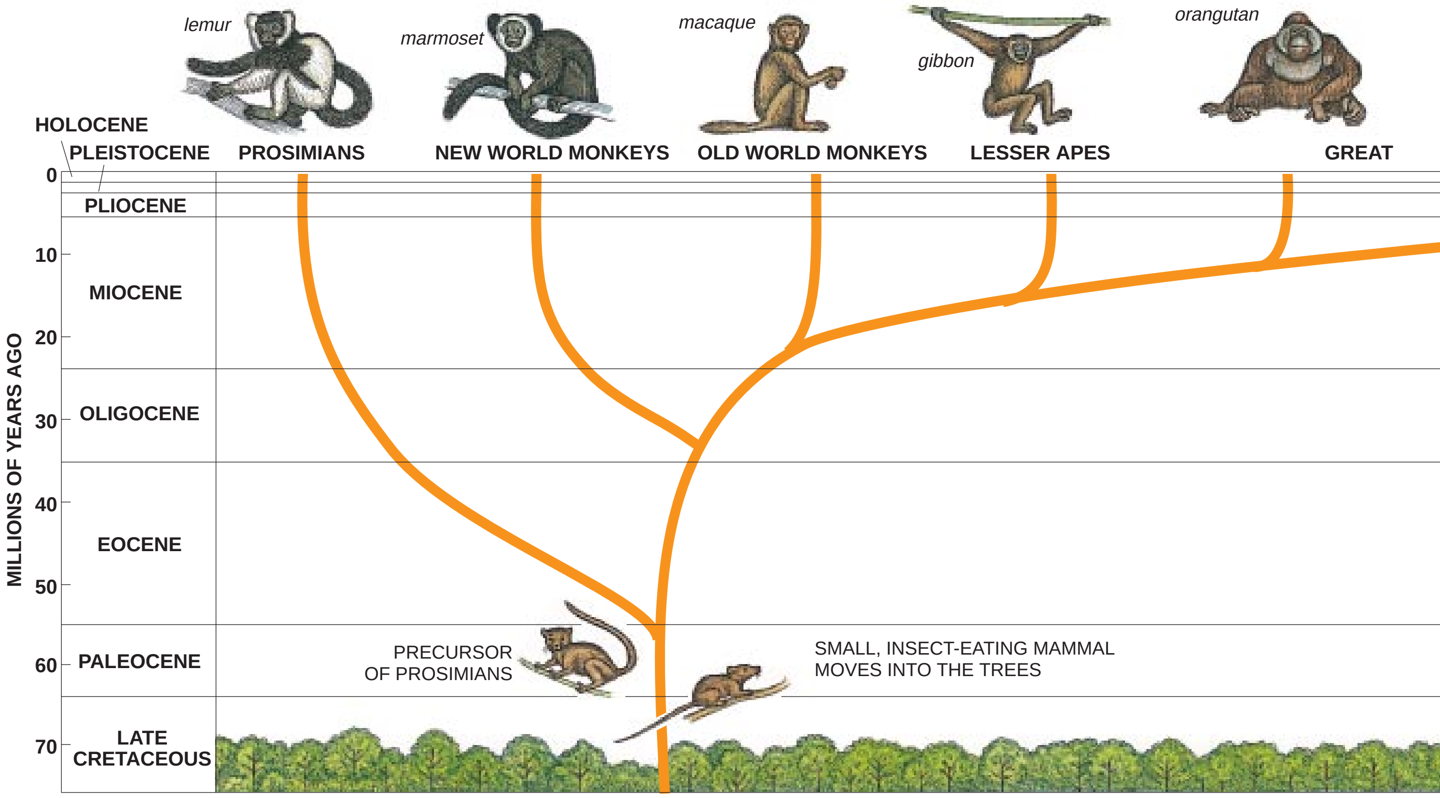
The dietary challenges trees and vines pose do not end there. Many plant foods lack one or more nutrients required by animals, such as particular vitamins or amino acids (the building blocks of protein), or else they are low in readily digestible carbohydrates (starch and sugar), which provide glucose and therefore energy. Usually, then, animals that depend primarily on plants for meeting their daily nutritional requirements must seek out a variety of complementary nutrient sources, a demand that greatly complicates food gathering.

For instance, most arboreal primates focus on ripe fruits and leaves, often supplementing their mostly herbivorous intake with insects and other animal matter. Fruits tend to be of high quality (rich in easily digested forms of carbohydrate and relatively low in fiber), but they provide little protein. Because all animals need a minimal amount of protein to function, fruit eaters must find additional sources of amino acids. Furthermore, the highest-quality items in the forest tend to be the most scarce. Leaves offer more protein and are more plentiful than fruit, but they are of lower quality (lower in energy and higher in fiber) and are more likely to include undesirable chemicals.

The need to mix and match plant foods is further exacerbated by the large distance between trees of the same species in tropical forests, which include hundreds of tree species. An animal that concentrated on eating food from a single species would have to exert great effort going from one individual of that species to another. What is more, trees exhibit seasonal peaks and valleys in the production of the fruits and young leaves primates like to eat, again making reliance on a single food species untenable.

From an evolutionary perspective, two basic strategies for coping with these many problems are open to a nascent plant eater. In one, morphology reigns supreme: over long time spans, natural selection may favor the acquisition of anatomic specializations—especially of the digestive tract—that ease the need to invest time and energy searching for only the highest-quality dietary items. That is, morphological adaptations enable animals to

YOUNG CHIMPANZEES SEEK FRUIT as part of a diet that consists primarily of ripe fruits supplemented by leaves and some animal prey. Obtaining the foods needed for adequate nutrition in the tropical forest turns out to be significantly more difficult for primates than was once believed. The author contends that the solutions adopted by primates millions of years ago strongly influenced the subsequent evolution of the primate order. The drawings on the opposite page depict some typical plant foods available to arboreal animals in the tropical forest.



EVOLUTIONARY TREE of the primate order is rooted in the late Cretaceous, when a small, insect-eating mammal climbed into the trees to take advantage of feeding opportunities presented by the spread of angiosperm forests. As the descendants of this mammal (*artist's representation to left of tree*) adapted to a new dietary niche in the canopy, they developed traits now regarded as characteristic of primates, such as a rounded snout and nails (instead of claws). These descendants

depend on plant parts that are ubiquitous, such as on mature leaves (which are readily available but not of particularly high quality).

Colobine monkeys, one of the Old World primate groups in Africa and Asia, offer an excellent example of this strategy. Unlike the typical primate digestive tract (including that of humans), with its simple acid stomach, that of colobines includes a compartmentalized, or sacculated, stomach functionally analogous to that of cows and other ruminants. This anatomic specialization enables colobines to process fiber extremely efficiently.

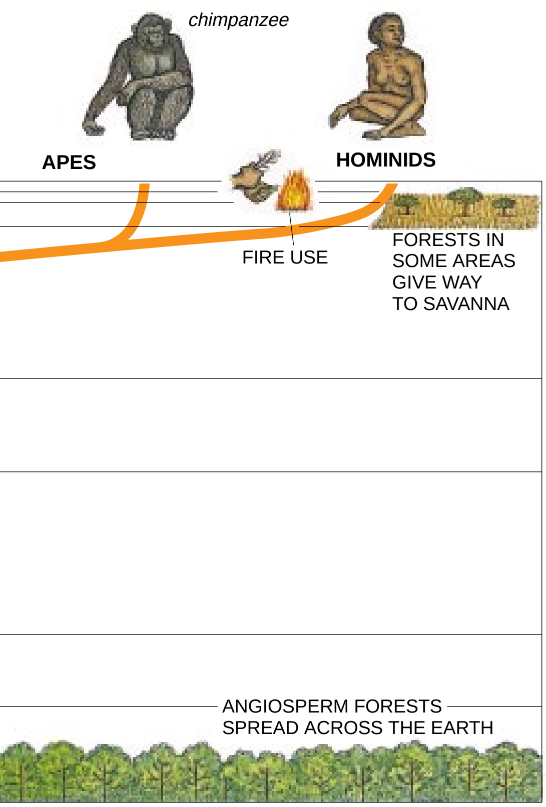
Chewed leaves flow through the esophagus into the forestomach, one of the two stomach compartments in colobines. In this alkaline forestomach, microbes known as cellulolytic bacteria do what digestive enzymes of the monkeys cannot do: degrade fiber. In a process known as fermentation, the bacteria break down the cellulose and hemicellulose in plant cell walls, using those substances as an energy source to fuel their own activities. As the bacteria consume the fiber, they release gases called volatile fatty acids. These gases pass through the stomach wall into the colobine bloodstream, where they provide energy for body tissues or are delivered to the liver for conversion into glucose. Some researchers think the colobine forestomach may also aid in the detoxification of harmful secondary compounds in plant foods.

Efficiency of nutrient extraction from fibrous foods is enhanced in another way in colobine monkeys. As cellulolytic bacteria die, they pass out of the forestomach into the second compartment, a simple acid stomach similar to our own. Here special enzymes (lysozymes) cleave the bacterial cell walls. In consequence, protein and other nutritious materials that compose the cellulolytic bacteria become available for digestion by the monkeys. (In a sense, then, once leaves are chewed and swallowed, colobine monkeys do not interact directly with their food; they live on products of the fermentation process and on the nutrients provided by the fermenters.)

In contrast to colobines, humans and most other primates pass fiber basically unchanged through their acid stomach and their small intestine (where most nutrients are absorbed) and into the hindgut (the cecum and colon). Once fiber reaches the hindgut, cellulolytic bacteria may be able to degrade some of it. But, for most primates, eating copious amounts of fiber does not confer the same benefits as it does for the digestively specialized colobines.

Another morphological change that can facilitate survival on lower-quality plant parts is to grow larger over time. Compared with small animals, big ones must consume greater absolute amounts of food to nourish their more extensive tissue mass. But, for reasons that are imperfectly understood, the bigger animals can actually attain adequate nourishment by taking in less energy per unit of body mass. This relatively lower energy demand means larger animals can meet their energy requirements with lower-quality foods. Growing bigger has been only a limited option for most primates, however. If arboreal animals grow too massive, they risk breaking the branches underneath their feet and falling to the ground.

The second basic strategy open to plant eaters is more behavioral than morphological. Species can opt to feed selectively on only the highest-quality plant foods. But because quality items are rare and very patchily distributed in tropical forests, this strategy requires the adoption of behaviors that help to minimize the costs of pro-



gave way to true primates, beginning with the prosimians. Our own genus, *Homo,* emerged during the Pliocene. Exact dates of radiations are debatable.

curing these resources. The strategy would be greatly enhanced by a good memory. For example, an ability to remember the exact locations of trees that produce desirable fruits and to recall the shortest routes to those trees would enhance foraging efficiency by lowering search and travel costs. So would knowledge of when these trees were likely to bear ripe fruits. Reliance on memory, with its attendant benefits, might then select for bigger brains having more area for storing information.

Of course, these two basic evolutionary strategies—the morphological and behavioral—are not mutually exclusive, and species vary in the extent to which they favor one or the other. As a group, however, primates have generally depended most strongly on selective feeding and on having the brain size, and thus the wit, to carry off this strategy successfully. Other plant-eating orders, in contrast, have tended to focus heavily on morphological adaptations.

I gained my first insights into the evolutionary consequences of selective feeding in primates in the mid-1970s, when I noticed that howler monkeys and black-handed spider monkeys (*Ateles geoffroyi*)—two New World primate species—favored markedly different diets. Howler and spider monkeys, which diverged from a common ancestor, are alike in that they are about the same size, have a simple, unsacculated stomach, are totally arboreal and eat an almost exclusively plant-based diet, consisting for the most part of fruits and leaves. But my fieldwork showed that the foundation of the howler diet in the Barro Colorado forest was immature leaves, whereas the foundation of the spider monkey diet was ripe fruits.

Most of the year howlers divided their daily feeding time about equally between new leaves and fruits. But during seasonal low points in overall fruit availability, they ate virtually nothing but leaves. In contrast, spider monkeys consumed ripe fruits most of the year, eating only small amounts of leaves. When fruits became scarce, spider monkeys did not simply fill up on leaves as the howlers did. Their leaf intake did increase, but they nonetheless managed to include considerable quantities of fruit in the diet. They succeeded by carefully seeking out all fruit sources in the forest; they even resorted to consuming palm nuts that had not yet ripened.

These observations raised a number of questions. I wanted to know how howlers obtained enough energy during months when they lived exclusively on leaves. As already discussed, much of the energy in leaves is bound up in fiber that is inaccessible to the digestive enzymes of primates. Further, why did howlers eat considerable foliage even when they had abundant access to ripe fruits? By the same token, why did spider monkeys go out of their way to find fruit during periods of scarcity; what stopped them from simply switching to leaves, as howlers did? And how did spider monkeys meet daily protein needs with their fruit-rich diet? (Recall that fruits are a poor source of protein.)

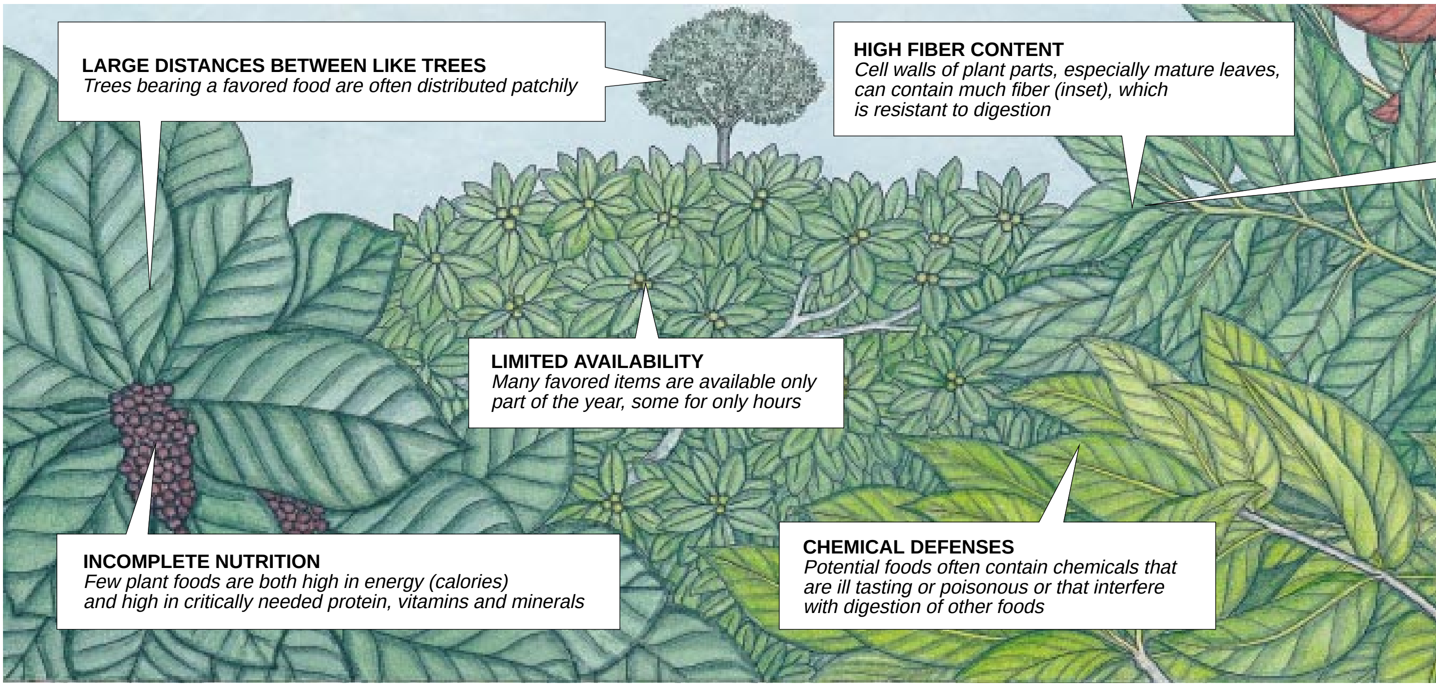
Because howler and spider monkeys are much alike externally, I speculated that some internal feature of the two species—perhaps the structure of the gut or the efficiency of digestion—might be influencing these behaviors. And, indeed, studies in which I fed fruits and leaves to temporarily caged subjects revealed that howler monkeys digested food more slowly than did spider monkeys. Howlers began eliminating colored plastic markers embedded in foods an average of 20 hours after eating. In contrast, spider monkeys began eliminating these harmless markers after only four hours. Examining the size of the digestive tract in the two species then revealed how these different passage rates were attained. In howler monkeys the colon was considerably wider and longer than in spider monkeys, which meant food had a longer distance to travel and that significantly more bulk could be retained.

Collectively, these results implied that howlers could survive on leaves because they were more adept at fermenting fiber in the cecum and colon. They processed food slowly, which gave bacteria in the capacious hindgut a chance to produce volatile fatty acids in quantity. Experiments I later carried out with Richard McBee of Montana State University confirmed that howlers may obtain as much as 31 percent of their required daily energy from volatile fatty acids produced during fermentation.

In contrast, spider monkeys, by passing food more quickly through their shorter, narrower colons, were less efficient at extracting energy from the fiber in their diet. This speed, however, enabled them to move masses of food through the gastrointestinal tract each day. By choosing fruits, which are highly digestible and rich in energy, they attained all the calories they needed and some of the protein. They then supplemented their basic fruit-pulp diet with a few very select young leaves that supplied the rest of the protein they required, without an excess of fiber.

Hence, howler monkeys never devote themselves exclusively to fruit, in part because their slow passage rates would probably prevent them from processing all the fruit they would need to meet their daily energy requirement. And the amount of fruit they could consume certainly would not provide enough protein. Conversely, spider monkeys must eat fruit because their digestive tract is ill equipped to provide great amounts of energy from fermenting leaves; efficient fermentation requires that plant matter be held in the gut for some time.

By luck, I had chosen to study two species that fell at opposite ends of the continuum between slow and rapid passage of food. It is now clear that most primate species can be ranked somewhere along this continuum, depending on whether they tend to maximize the efficiency with which they digest a given meal or maximize the volume of food processed in a day. This research further shows that even without major changes in the design of the digestive tract, subtle adjustments in the size of different segments of the gut can help compensate for nutritional problems posed by an animal's dietary choices. Morphological compensations in the digestive tract can have their drawbacks, however, because they may make it difficult for a species to alter its dietary habits should environmental conditions change suddenly.



These digestive findings fascinated me, but a comparison of brain size in the two species yielded one of those "eurekas" of which every scientist dreams. I examined information on the brain sizes of howler and spider monkeys because the spider monkeys in Panama seemed "smarter" than the howlers—almost human. Actually, some of them reminded me of my friends. I began to wonder whether spider monkeys behaved differently because their brains were more like our own. My investigations showed that, indeed, the brains of howler and spider monkeys do differ, even though the animals are about the same size. (Same-sized animals generally have like-sized brains.) The spider monkey brain weighs about twice that of howlers.

Now, the brain is an expensive organ to maintain; it usurps a disproportionate amount of the energy (glucose) extracted from food. So I knew natural selection would not have favored development of a large brain in spider monkeys unless the animals gained a rather pronounced benefit from the enlargement. Considering that the most striking difference between howler and spider monkeys is their diets, I proposed that the bigger brain of spider monkeys may have been favored because it facilitated the development of mental skills that enhanced success in maintaining a diet centered on ripe fruit.

A large brain would certainly have helped spider monkeys to learn and, most important, to remember, where certain patchily distributed fruit-bearing trees were located and when the fruit would be ready to eat. Also, spider monkeys comb the forest for fruit by dividing into small, changeable groups. Expanded mental capacity would have helped them to recognize members of their particular social unit and to learn the meaning of the different food-related calls through which troop members convey over large distances news of palatable items. Howler monkeys, in contrast, would not need such an extensive memory, nor would they need so complex a recognition and communication system. They forage for food as a cohesive social unit, following well-known arboreal pathways over a much smaller home range.

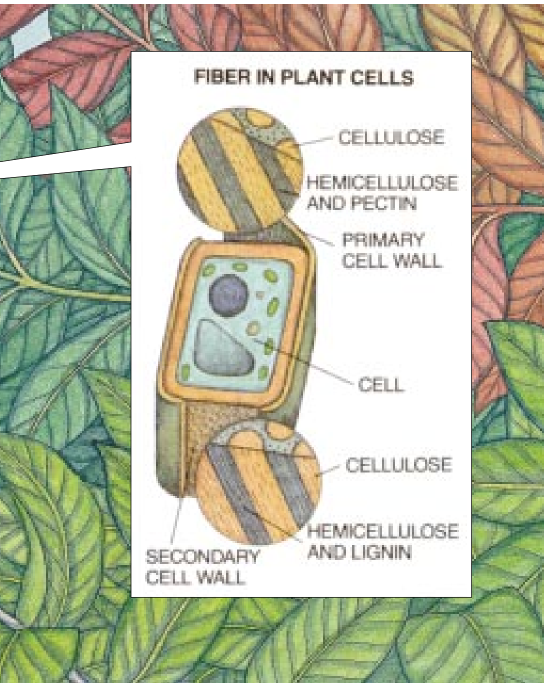
If I was correct that the pressure to obtain relatively difficult-to-find, high-quality plant foods encourages the development of mental complexity (which is paid for by greater foraging efficiency), I would expect to find similar differences in brain size in other primates. That is, monkeys and apes who concentrated on ripe fruits would have larger brains than those of their leaf-eating counterparts of equal body size. To pursue this idea, I turned to estimates of comparative brain sizes published by Harry J. Jerison of the University of California at Los Angeles. To my excitement, I found that those primate species that eat higher-quality, more widely dispersed foods generally have a larger brain than do their similar-sized counterparts that feed on lower-quality, more uniformly distributed resources.

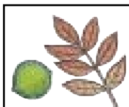
As I noted earlier, primates typically have larger brains than do other mammals of their size. I believe the difference arose because primates feed very selectively, favoring the highest-quality plant parts—for instance, even primates that eat leaves tend to choose very immature leaves or only the low-fiber tips of those leaves.

Having uncovered these links between dietary pressures and evolution in nonhuman primates, I became curious about the role of such pressures in human evolution. A review of the fossil record for the hominid family—humans and their precursors—provided some intriguing clues.

Australopithecus, the first genus in our family, emerged in Africa more than 4.5 million years ago, during the Pliocene. As is true of later hominids, they were bipedal, but their brains were not appreciably larger than those of today's apes. Hence, selection had not yet begun to favor a greatly enlarged brain in our family. The fossil record also indicates *Australopithecus* had massive molar teeth that would have been well suited to a diet consisting largely of tough plant material. Toward the end of the Pliocene, climate conditions began to change. The next epoch, the Pleistocene (lasting from about two million to 10,000 years ago), was marked by repeated glaciations of the Northern Hemisphere. Over both epochs, tropical forests shrank and were replaced in many areas by savanna woodlands.

As the diversity of tree species decreased and the climate became more seasonal, primates in the expanding savanna areas must have faced many new dietary challenges. In the Pleistocene the last species of *Australopithecus*—which



	READILY ACCESSIBLE CALORIES	PROTEIN	FIBER	CHEMICAL DEFENSES	AVAILABILITY ON A GIVEN TREE
FLOWERS	Moderate	Moderate to high	Low to moderate	Variable	Fewer than three months
FRUITS	High	Low	Moderate	Low	Fewer than three months
YOUNG LEAVES	Low	High	Moderate	Moderate	Half the year
MATURE LEAVES	Low	Moderate	High	Moderate	Almost year-round

MANY CHALLENGES can deter primates in the tropical forest from obtaining the calories and mix of nutrients they need from plant foods (*left*). Because most such foods are inadequate in one way or another, animals must choose a variety of items each day. The chart at the right loosely reflects the relative abundance of desirable (*green*) and problematic (*yellow*) components in a mouthful of common foods. It also indicates the typical availability of these foods on any given tree.

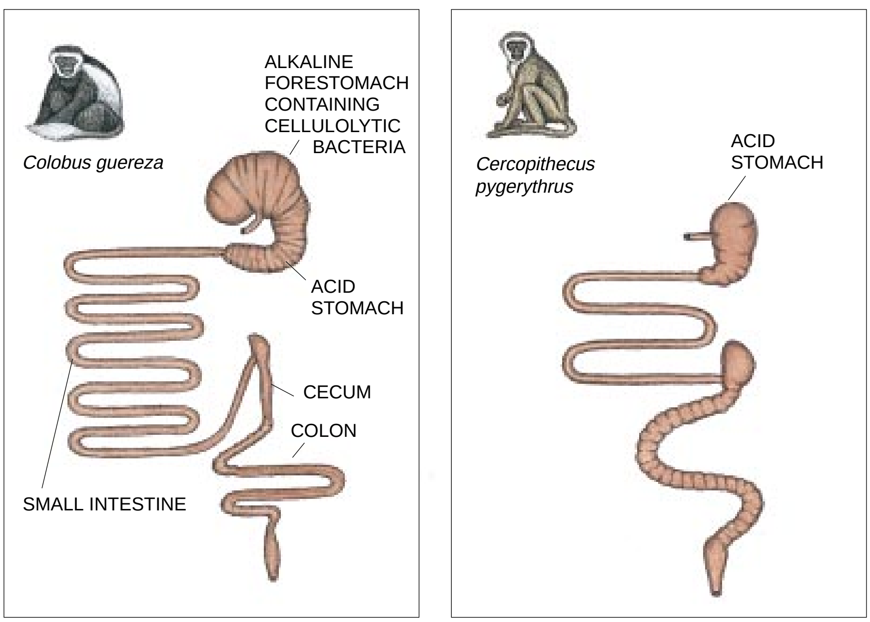
by then had truly massive jaws and molars—went extinct. Perhaps those species did so, as my colleague Montague W. Demment of the University of California at Davis speculates, because they were outcompeted by the digestively specialized ungulates (hoofed animals).

The human, or *Homo,* genus emerged during the Pliocene. The first species of the genus, *H. habilis,* was similar in body size to *Australopithecus* but had a notably larger brain. This species was replaced by the even larger-brained *H. erectus* and then, in the Pleistocene, by *H. sapiens,* which has the biggest brain of all. In parallel with the increases in brain size in the *Homo* genus, other anatomic changes were also occurring. The molar and premolar teeth became smaller, and stature increased.

To me, the striking expansion of brain size in our genus indicates that we became so successful because selection amplified a tendency inherent in the primate order since its inception: that of using brain power, or behavior, to solve dietary problems. Coupled with the anatomic changes—and with the associations in living primates between larger brains and a high-quality diet—this increase also points to the conclusion that the behavioral solution was to concentrate on high-quality foods. In fact, I suspect early humans not only maintained dietary quality in the face of changing environmental conditions but even improved it.

Expansion of the brain in combination with growth in body size and a reduction in the dentition supports the notion of a high-quality diet for a couple of reasons. When one examines present-day orangutans and gorillas, it becomes clear that in our superfamily, Hominoidea (apes and humans), an increase in body size combined with *decreased* dietary quality leads to a slow-moving, fairly sedentary and unsociable ape. Yet our *Homo* ancestors apparently were mobile and sociable—more resembling the lively, social and communicative chimpanzee. Unlike orangutans and gorillas, chimpanzees feed preferentially on high-quality, energy-rich ripe fruits.

Likewise, the reduction in the molars and premolars shows that the texture of foods we ate had somehow been altered such that the dentition no longer had so much work to do. In other words, either these early humans were eating differ-



DIGESTIVE TRACT of colobine monkeys, such as that in *Colobus guereza* (*left*), is specialized: the stomach consists of two distinct compartments instead of the single chamber found in vervet monkeys (*right*) and most other primates. One of those compartments—the forestomach—is designed to extract more energy from fiber than would normally be obtainable. Colobine monkeys can thus survive on a more fibrous diet than can other primates of similar size.

ent (less fibrous, easier-to-chew) foods than was *Australopithecus,* or they were somehow processing foods to remove material that would be hard to chew and digest. Indeed, stone tools found with fossil remains of *H. habilis* indicate that even the earliest members of our genus were turning to technology to aid in the preparation of dietary items.

The probability that hominids persisted in seeking energy-rich foods throughout their evolution suggests an interesting scenario. As obtaining certain types of plant foods presumably became more problematic, early humans are thought to have turned increasingly to meat to satisfy their protein demands. One can readily envision their using sharp stone flakes to cut through tough hides and to break bones for marrow. To incorporate meat into the diet on a steady basis and also to amass energy-rich plant foods, our ancestors eventually developed a truly novel dietary approach. They adopted a division of labor, in which some individuals specialized in the acquisition of meat by hunting or scavenging and other individuals specialized in gathering plants. The foods thus acquired were saved instead of being eaten on the spot; they were later shared among the entire social unit to assure all members of a balanced diet.

Survival of the individual thus came to depend on a number of technological and social skills. It demanded not only having a brain able to form and retain a mental map of plant food supplies but also having knowledge of how to procure or transform such supplies. In addition, survival now required an ability to recognize that a stone tool could be fashioned from a piece of a rock and a sense of how to implement that vision. And it required the capacity to cooperate with others (for instance, to communicate about who should run ahead of a hunted zebra and who behind), to defer gratification (to save food until it could be brought to an agreed site for all to share) and both to determine one's fair portion and to ensure that it was received. Such demands undoubtedly served as selective pressures favoring the evolution of even larger, more complex brains.

Similarly, spoken communication may at first have helped facilitate the cooperation needed for efficient foraging and other essential tasks. Gradually, it became elaborated to smooth the course of social interactions.

In other words, I see the emergence and evolution of the human line as stemming initially from pressures to acquire a steady and dependable supply of very high quality foods under environmental conditions in which new dietary challenges made former foraging behaviors somehow inadequate. Specialized carnivores and herbivores that abound in the African savannas were evolving at the same time as early humans, perhaps forcing them to become a new type of omnivore, one ultimately dependent on social and technological innovation and thus, to a great extent, on brain power. Edward O. Wilson of Harvard University has estimated that for more than two million years (until about 250,000 years ago), the human brain grew by about a tablespoon every 100,000 years. Apparently each tablespoonful of brain matter added in the genus *Homo* brought rewards that favored intensification of the trend toward social and technological advancement.

Although the practice of adding some amount of meat to the regular daily intake became a pivotal force in the emergence of modern humans, this behavior does not mean that people today are biologically suited to the virtually fiber-free diet many of us now consume. In fact, in its general form, our digestive tract does not seem to be greatly modified from that of the common ancestor of apes and humans, which was undoubtedly a strongly herbivorous animal.

Yet as of the mid-1980s no studies had been done to find out whether the gut functions of modern humans were in fact similar to those of apes. It was possible that some functional differenc-

SPIDER MONKEY
(*Ateles geoffroyl*)

TYPICAL DIET
Fruits: *72 percent*
Leaves: *22 percent*
Flowers: *6 percent*

WEIGHT
Six to eight kilograms

BRAIN SIZE
107 grams

DAY RANGE
915 meters

DIGESTIVE FEATURES
Small colon
Fast passage of food through colon

HOWLER MONKEY
(*Alouatta palliata*)

TYPICAL DIET
Fruits: *42 percent*
Leaves: *48 percent*
Flowers: *10 percent*

WEIGHT
Six to eight kilograms

BRAIN SIZE
50.3 grams

DAY RANGE
443 meters

DIGESTIVE FEATURES
Large colon
Slow passage of food through colon

SPIDER MONKEY (*left*) is a fruit specialist, whereas the howler monkey (*right*) eats large quantities of leaves. The author proposes that diet played a major role in shaping the different traits of the two like-sized species, which shared a common ancestor. Natural selection favored a larger brain in spider monkeys, in part because enhanced mental capacity helped them remember where ripe fruits could be found. And spider monkeys range farther each day because in any patch of forest, ripe fruits are less abundant than leaves. The digestive traits of spider and howler monkeys promote efficient extraction of nutrition from fruits and leaves, respectively.

es existed, because anatomic evidence had shown that despite similarity in the overall form of the digestive tract, modern humans have a rather small tract for an animal of their size. They also differ from apes in that the small intestine accounts for the greatest fraction of the volume of the human digestive tract; in apes the colon accounts for the greatest volume.

To better understand the kind of diet for which the human gut was adapted, Demment and I decided to compare human digestive processes with those of the chimpanzee, our closest living relative. We hoped to determine whether, over the course of their respective evolutionary histories, humans and chimpanzees had diverged notably in their abilities to deal with fiber. (We were greatly encouraged in this effort by the late Glynn Isaac, who was then at the University of California at Berkeley.)

The feeding habits of chimpanzees are well known. Despite their skill in capturing live prey (particularly monkeys), these apes actually obtain an estimated 94 percent of their annual diet from plants, primarily ripe fruits. Even though the fruits chimpanzees eat tend to be rich in sugar, they contain far less pulp and considerably more fiber and seeds than do the domesticated fruits sold in our supermarkets. Hence, I calculated that wild chimpanzees take in hundreds of grams of fiber each day, much more than the 10 grams or less the average American is estimated to consume.

Various excellent studies, including a fiber project at Cornell University, had already provided much information about fiber digestion by humans. At one time, it was believed that the human digestive tract did not possess microbes capable of degrading fiber. Yet bacteria in the colons of 24 male college students at Cornell proved quite efficient at fermenting fiber found in a variety of fruits and vegetables. At their most effective, the microbial populations broke down as much as three quarters of the cell-wall material that the subjects ingested; about 90 percent of the volatile fatty acids that resulted were delivered to the bloodstream.

Following the example of the Cornell study, Demment and I assessed the efficiency of fiber breakdown in chimpanzees fed nutritious diets containing varying amounts of fiber. Demment handled the statistical analyses, and I collected raw data. How dry that sounds in comparison to the reality of the experience! At the Yerkes Primate Center in Atlanta, I whiled away the summer with six extremely cross chimpanzees that never missed an opportunity to pull my hair, throw fecal matter and generally let me know they were underwhelmed by our experimental cuisine.

BURGER AND FRIES, like many popular foods eaten in the U.S., bear little resemblance to the fruits and leaves most primates have emphasized since the inception of our order. Early humans, too, are thought to have consumed large quantities of plant foods. Hence, modern diets often diverge greatly from those to which the human body may be adapted.

Our results showed that the chimpanzee gut is strikingly similar to the human gut in the efficiency with which it processes fiber. Moreover, as the fraction of fiber in the diet rises (as would occur in the wild during seasonal lulls in the production of fruits or immature leaves), chimpanzees and humans speed the rate at which they pass food through the digestive tract.

These similarities indicate that as quality begins to decline in the natural environment, humans and chimpanzees are evolutionarily programmed to respond to this decrease by increasing the rate at which food moves through the tract. This response permits a greater quantity of food to be processed in a given unit of time; in so doing, it enables the feeder to make up for reduced quality by taking in a larger volume of food each day. (Medical research has uncovered another benefit of fast passage. By speeding the flow of food through the gut, fiber seems to prevent carcinogens from lurking in the colon so long that they cause problems.)

If the human digestive tract is indeed adapted to a plant-rich, fibrous diet, then this discovery lends added credence to the commonly heard assertion that people in highly technological societies eat too much refined carbohydrate and too little fiber. My work offers no prescription for how much fiber we need. But certainly the small amount many of us consume is far less than was ingested by our closest human ancestors.

More recently, my colleagues and I have analyzed plant parts routinely eaten by wild primates for their content of various constituents, including vitamin C and pectin. Pectin, a highly fermentable component of cell walls, is thought to have health benefits for humans. Our results suggest that diets eaten by early humans were extremely rich in vitamin C and contained notable pectin. Again, I do not know whether we need to take in the same proportions of these substances as wild primates do, but these discoveries are provocative.

To a major extent, the emergence of modern humans occurred because natural selection favored adaptations in our order that permitted primates to focus their feeding on the most energy-dense, low-fiber diets they could find. It seems ironic that our lineage, which in the past benefited from assiduously avoiding eating too much food high in fiber, may now be suffering because we do not eat enough of it.

FURTHER READING

ECOLOGY OF ARBOREAL FOLIVORES. Edited by G. Gene Montgomery. Smithsonian Institution Press, 1978.

DISTRIBUTION PATTERNS OF TROPICAL PLANT FOODS AS AN EVOLUTIONARY STIMULUS TO PRIMATE MENTAL DEVELOPMENT. K. Milton in *American Anthropologist,* Vol. 83, No. 3, pages 534–548; September 1981.

FOOD CHOICE AND DIGESTIVE STRATEGIES OF TWO SYMPATRIC PRIMATE SPECIES. K. Milton in *American Naturalist,* Vol. 117, No. 4, pages 496–505; April 1981.

PRIMATE DIETS AND GUT MORPHOLOGY: IMPLICATIONS FOR HOMINID EVOLUTION. K. Milton in *Food and Evolution: Toward a Theory of Human Food Habits.* Edited by Marvin Harris and Eric B. Ross. Temple University Press, 1987.

DIGESTION AND PASSAGE KINETICS OF CHIMPANZEES FED HIGH AND LOW-FIBER DIETS AND COMPARISON WITH HUMAN DATA. K. Milton and M. W. Demment in *Journal of Nutrition,* Vol. 118, No. 9, pages 1082–1088; September 1988.

FORAGING BEHAVIOUR AND THE EVOLUTION OF PRIMATE INTELLIGENCE. K. Milton in *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Edited by Richard Byrne and Andrew Whiten. Oxford University Press, 1988.

The Great Radium Scandal

William J. A. Bailey grew rich from his radium-laced patent medicine until it killed a leading socialite. The scandal helped to usher in modern standards of radioisotope regulation

by Roger M. Macklis

Eben M. Byers died early on the morning of Thursday, March 31, 1932, the victim of a mysterious syndrome that for 18 months had ravaged his body, corroding his skeletal system until one by one his bones started to splinter and break. Byers had been a powerful man, a broad-chested athlete and sportsman who was an expert trapshooter and had been the U.S. Amateur Golf Champion in 1907 at the age of 27. As chairman of the A. M. Byers Iron Foundry, he had personified the Roaring Twenties, a millionaire socialite and tycoon who had clambered into the upper reaches of New York society. He continued to lead a life of privilege even after the stock market crash, maintaining homes in Pittsburgh, New York, Rhode Island and South Carolina, as well as horse-racing stables in New York and England.

When Byers died, his shriveled body must have been barely recognizable to friends who had known him as a robust athlete and ladies' man. He weighed just 92 pounds. His face, once youthful and raffishly handsome, set off by dark, pomaded hair and deep-set eyes, had been disfigured by a series of last-ditch operations that had removed most of his jaw and part of his skull in a vain attempt to stop the destruction of bone. His marrow and kidneys had failed, giving his skin a sallow, ghastly cast. Although a brain abscess had left him nearly mute, he remained lucid almost to the end. He died at 7:30 A.M. at Doctors' Hospital in New York City.

News of Byers's death and its mysterious circumstances reached his former colleagues on Wall Street almost immediately. Over the next two weeks the stock of his company, already battered by the Great Depression, lost one third of its value. Worried friends and relatives had begun to contact Byers's doctors from the day of his death to find out if he had died of something contagious. By the next afternoon, the authorities had begun a criminal poisoning investigation and were preparing the body for a forensic autopsy by the chief medical examiner of New York. The *New York Times* announced the preliminary results in a front-page headline: "Eben M. Byers Dies Of Radium Poisoning."

Radium poisoning? How could a man of Byers's position and wealth have suffered a malady that so far had been confined to a handful of radium chemists and dial painters who used radioactive ink to make watch faces glow in the dark, licking their brushes to draw a finer line? The answer to this question focused attention on a danger that public health officials had only just begun to recognize: even small amounts of toxic substances can kill, over time. The Byers case thus helped to create the presumption that medicines are dangerous until proved safe.

Byers's radioactive saga had started on a chartered train returning from the Harvard-Yale game of 1927. Engaging in some late-night revelry in a private Pullman booth, Byers fell from his berth and injured his arm. Despite the best ministrations of his personal physicians and trainers, Byers complained that the ache simply would not go away. Soon the injury was affecting his golf game (and, it was rumored, his libido). Even-

ROGER M. MACKLIS is chairman of the department of radiation oncology at the Cleveland Clinic Foundation. He divides his time between his clinical practice in pediatric radiation oncology and his research into novel anticancer radiopharmaceuticals. He has written on the history of medicine in a number of professional journals. He is now preparing a section on radiomedical quackery in the early 1900s for a series to be published in 1994 under the aegis of the American College of Radiology in honor of the 100th anniversary of the discovery of x-rays. Macklis holds a master's degree in biophysics from Yale University and a degree from Harvard Medical School.

IMMODEST PROPOSAL from the Roaring Twenties promotes Radithor, a radium-laced patent medicine, and its purveyor, William J. A. Bailey, a college dropout who gave himself the title of doctor. Such advertisements capitalized on the fad for radium, whose efficacy in destroying tumors made it seem a panacea.

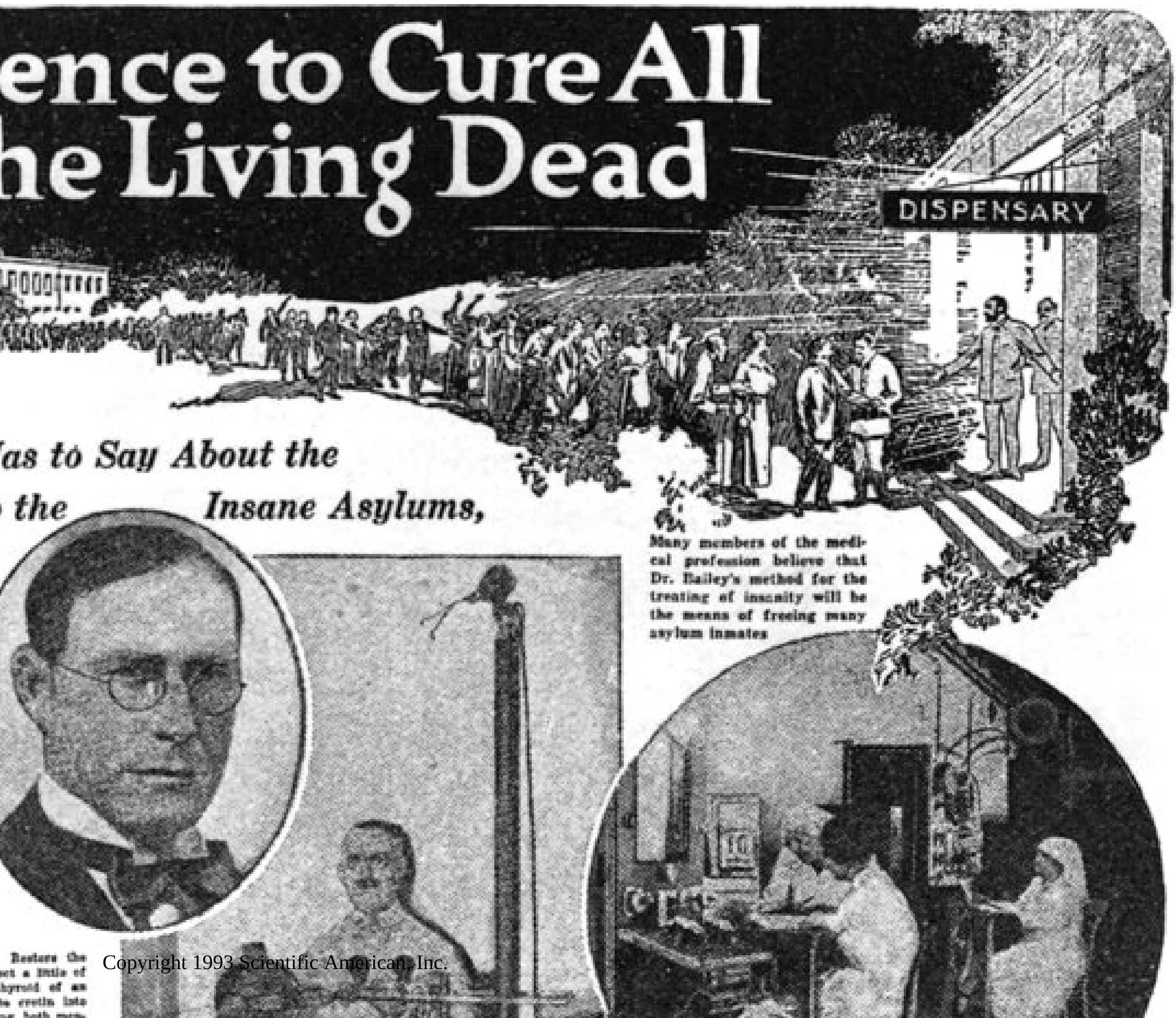
tually Byers found his way to a Pittsburgh physician named Charles Clinton Moyar, who suggested that Byers try Radithor, a patent medicine whose manufacturer, the Bailey Radium Laboratory in New Jersey, described it as a cure for dyspepsia, high blood pressure, impotence and more than 150 other "endocrinologic" maladies.

Byers began drinking several bottles a day beginning in December 1927. He told friends that he felt invigorated and rejuvenated. So satisfied was he with the results that he sent cases to his friends, colleagues and female acquaintances and even fed some of the expensive potion to his racehorses. Like other ardent Radithor enthusiasts, Byers apparently consumed vast quantities of the nostrum, drinking between 1,000 and 1,500 bottles between 1927 and 1931. He probably accumulated a radiation dosage equivalent to thousands of x-rays—perhaps three times the lethal dose, if absorbed all at once.

The autopsy conducted the day after Byers died confirmed that his bones and organs were dangerously radioactive. Placed on a film-plate overnight in the dark, his extracted teeth and jawbone produced a dramatic film exposure pattern. As the evidence for radium poisoning grew, squads of public health officers pulled the dangerous materials off store shelves. Nervous citizens sheepishly came forward to turn in their own radioactive medicines. Among them was Mayor James J. Walker of New York City, who at first refused to give up his radioactive rejuvenator because, he said, it made him feel so good.

My own involvement with the strange case of Eben Byers began in the fall of 1989, when I came across several empty bottles of Radithor in a medical antiques shop and bought one on a whim. Because my laboratory research centers on treating cancer with biologically targeted radioactive compounds, I knew it was possible to make water temporarily radioactive by incubating it with radium. The radium gives off radon, a radioactive gas whose half-life is short. I assumed that the maker of the patent medicine had resorted to this very inexpensive process and that the Radithor's residual activity had decayed to insignificance long ago.

I was wrong. Tests performed by my colleagues John L. Humm and Marc R. Bellerive in our gamma-ray spectroscopy unit at the Dana-Farber Cancer Institute in Boston revealed that almost 70 years after it had been produced, the nearly empty bottle was still dangerously radioactive. We estimated that the original bottle must have contained approximately one microcurie each of radium 226 and radium 228. Intrigued, I bought the rest of the bottles. When they tested equally radioactive, I went to the Countway Medical Library rare book collection and the stacks of the Har-



vard University Library to find out more about the lost story of Radithor, the Bailey Radium Laboratory and its director and chief scientist, an inventor, entrepreneur and marketing genius named "Doctor" William J. A. Bailey.

Although contemporaries, William Bailey and Eben Byers came from opposite ends of the social spectrum. In a sense, they represented the twin faces of the American dream during the early decades of the 20th century. Byers was born in 1880 into a life of privilege, attending St. Paul's School and Yale University, where his suave demeanor and social conquests at nearby girls' schools earned him the nickname "Foxy Grandpa." After graduation in 1901, he dabbled in business, traveled and golfed, playing a major role in the Harvard-Yale defeat of the visiting Oxford-Cambridge golf team in 1903. The next year he became president and a director of the small Girard Iron Company, which had been built up by his father, Alexander Byers. In 1909, after the death of his older brother, Byers became president and eventually chairman of the iron foundry.

In contrast, William John Aloysius Bailey was born into a tough neighborhood in Boston on May 25, 1884. His father, a cook, died when he was young, and Bailey and his eight siblings were raised by their mother, Mary, on a weekly income of $15. Bailey attended Quincy Grammar School and graduated about 12th in his class from the prestigious Boston Public Latin School, long known as a stepping-stone to the Ivy League for poor boys with quick wits. He did poorly on his Harvard entrance examinations but managed to be admitted as a freshman in the fall of 1903. Mounting debts forced him to drop out two years later, and although Bailey later claimed to be a Harvard graduate with a doctorate from the University of Vienna, no evidence exists to support either claim.

Bailey moved to New York City, working in an import-export business and editing an export catalogue. His letters from the time speak enthusiastically of his master plan: to be appointed the unofficial U.S. trade ambassador to the imperial government of China. His scheme never materialized. Instead, in the years leading up to World War I, Bailey traveled widely, acquiring a cosmopolitan veneer that later served him well. The outbreak of war found him drilling oil in Russia, but he headed home to tinker in a mechanical workshop when wartime commerce proved impractical.

It is at this period that Bailey's name first began to appear in connection with various scams. On May 8, 1915, the *New York Times* reported that he had been arrested on charges of running a mail-order swindle out of the Carnegie Engineering Corporation (a paper firm with no relation to the Carnegie steel empire). He accepted advance mail deposits for a $600 car to be assembled in Michigan and delivered to a pickup point in Pittsburgh. No factory existed, and Bailey and two others were found guilty and sentenced to 30 days in jail.

In May of 1918, Bailey was fined $200 plus costs for fraudulently promoting

EBEN M. BYERS, athlete and ladies' man, as he appeared in his prime, several years before his consumption of Radithor led to the destruction of his bones and, in 1932, to his death. The scandal swept radioactive potions from drugstore shelves.

WILLIAM BAILEY had compiled a history of shady business dealings before striking gold with Radithor. He seems, however, to have believed his own propaganda, having taken his own medicine over a period of years.

a patent medicine for male impotence called Las-I-Go For Superb Manhood. Chemical analysis of these pills revealed that the active ingredient was strychnine. This episode appears to mark the beginning of Bailey's fascination with sexual stimulants and aphrodisiacs.

Radiation research was a natural draw for Bailey because it had become a glamour field in medicine. This glamour derived from the novelty of the phenomenon and its evident value in imaging the body and destroying tumors. Workers generally believed other, more subtle applications were waiting to be found. For instance, Marie Curie, who with her husband, Pierre, had discovered radium in 1898, set as one of her major postwar tasks the investigation of the effects of minute quantities of radium on cells, animals and humans. In England, this field was called mild radium therapy to differentiate it from the use of much larger doses in the treatment of cancer.

Mild radium therapy can be traced to the homeopathic and physical medicine theories of the 19th century. They held that most healing processes were natural and that tiny quantities of naturally occurring materials, coupled with exercise and sunlight, could cure most maladies. Proponents of these theories believed in the legendary healing powers of the great European hot springs. One mystery persisted, however. The waters appeared to lose their potency just a few days after they were bottled. In 1903 the discovery was made that the apparent pharmacological agent dissolved in these waters was radon. After Ernest Rutherford's investigation of alpha-particle emissions from radium and radon, the transient healing effects the hot springs were reputed to have were ascribed to these particles.

Workers hypothesized that the alpha particles might account for the operation of the endocrine system, a connection that is not as strange as it may seem. Both fields had recently come to fascinate the medical world, and each had won the highest honors. In 1921 Frederick Soddy received the Nobel Prize in Chemistry for his work on radioisotopes. In that same year, Frederick G. Banting and Charles H. Best isolated insulin, work for which Banting and John J. R. MacLeod won the 1923 Nobel Prize in Physiology or Medicine. The one discovery provided a new kind of energy, the other, a method of controlling the body's transduction of energy—the process by which sugars and other basic foods are converted into more readily usable forms. Might radioactivity be the spark that set the biophysical machinery in motion?

RADITHOR VIAL from a curio shop turned out to be dangerously radioactive some 70 years after its original purchaser had nearly drained it. This finding supports at least one of Bailey's claims: he actually did use radium, rather than a cheaper, short-lived isotope. The vial originally contained about a microcurie each of radium 226 and radium 228.

The German physiologist George Wendt, in his address to the 13th International Congress of Physiologists, reported that human leukocytes exposed to low-level radium radiation began migrating toward the radium source and that moribund vitamin-starved rats could temporarily be rejuvenated by exposure to radium. Like the stuff of homeopathic legends, radium appeared to be a substance with two distinct modes of medical efficacy: in large quantities, it was destructive, but in trace amounts, it was beneficent, perhaps even necessary.

Legends translated into products more quickly then than they do now. The feeble jurisdiction of the Food and Drug Administration did not extend to radium, which it classified as a natural element rather than a drug. Radioactive candies, liniments, potions and creams were widely available by 1915. At first, the fashion appears to have been for the most part confined to Europe. In the U.S., interest in the medicinal, "catalytic" properties of radium and its decay products surged after double Nobel laureate Marie Curie made a whistle-stop railway tour across the country in 1921.

It is not known whether William Bailey actually encountered Curie, but it is clear that beginning in the early 1920s he became enraptured with radioactivity and its effects on life. He produced a translation of Curie's 1910 classic, *Traité de Radioactivité*. He incorporated a company in New York City called Associated Radium Chemists, Inc., which put out a line of radioactive patent medicines, including Dax for coughs, Clax for influenza and Arium for rundown metabolisms. Before the Byers affair, this operation was closed by the Department of Agriculture on grounds of fraudulent advertising.

Bailey soon went on to found two new firms in New York City. The Thorone Company (Thorium Hormones) produced cure-alls containing radium and thorium for "all glandular, metabolism and faulty chemistry conditions," especially impotence. The American Endocrine Laboratory produced the Radioendocrinator, a gold-plated radium-containing harness that could be worn around the neck (to rejuvenate the thyroid), around the trunk (to irradiate the adrenals or ovaries) or, for enervated males, under the scrotum in a special jockstrap. The device sold first for $1,000, then $500 and then finally $150 as the market became saturated.

Bailey sought opportunities to present his theories at legitimate scientific meetings. In a public relations coup, he managed to secure an invitation to speak at the medicinal products session of the American Chemical Society meeting in Washington, D.C., in 1924. "We have cornered aberration, disease, old age, and in fact life and death themselves in the endocrines!" Bailey thundered. "In and around these glands must center all future efforts for human regeneration." The next day the *New York Times* excerpted the talk in a lengthy and complimentary article.

In 1925 Bailey moved to East Orange, N.J., and opened the Bailey Radium Laboratories. It was here that he created and bottled his promotional masterpiece, Radithor. Ironically, 1925 also marked the beginning of the end for mild radium therapy. A group of New Jersey radium chemists and dial painters working at the U.S. Radium Corporation died after a protracted and mysterious syndrome of kidney disease, low blood counts and widespread deterioration of their bones. Some experts muttered about possible radium poisoning, but Bailey disagreed. "There is no proof that radium was responsible for the deaths," he opined, when called by the *New York Times* for an expert comment.

His company shipped promotional pamphlets to every registered physician in the U.S., filling them with testimonials from patients and physicians as well as with photographs purporting to show the extraction, purification and testing of radium at the Bailey Labs. In fact, Bailey simply bought purified radium wholesale from the nearby American Radium Laboratory, bottling it in distilled water and marking up the price by almost 500 percent. He offered physicians a 17 percent rebate as a "professional fee," a practice that the American Medical Association condemned as "fee-splitting quackery" in 1927.

The promotions quickly made Bailey a rich man. He sold more than 400,000 half-ounce bottles between 1925 and 1930. Yet although evidence began to mount that small quantities of radioactive material could be devastating to health, the public at first took little notice. No one seemed to worry about a sickness that had so far been confined to poor, working-class women who painted radium onto watch dials. The FDA issued warnings, but it had no recourse to legal action. The Federal Trade Commission therefore took the lead by beginning an investigation of Bailey's claims in 1928. On February 5, 1930, the agency filed an official complaint charging Bailey with falsely advertising the efficacy and harmlessness of his products.

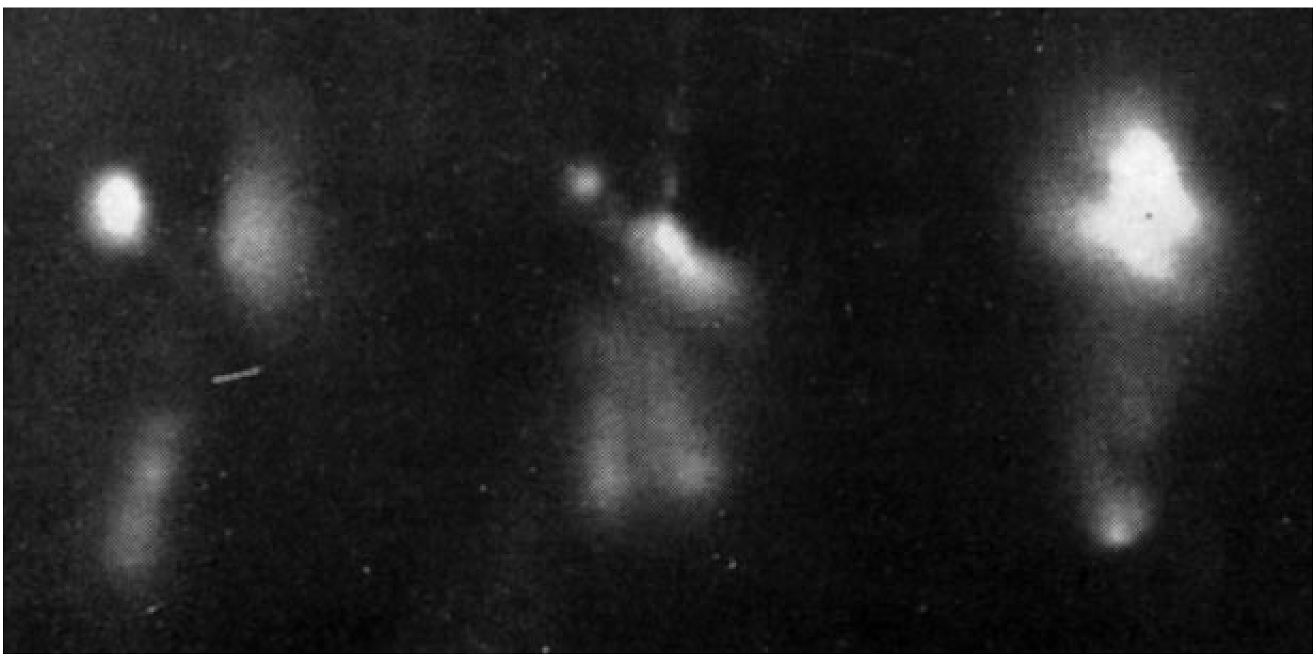
It was at about this time that Eben Byers began to experience unusual aches and pains. He told his private physician that he had lost "that toned-up feeling." He began to lose weight and complained of headaches and toothaches. He was told that he just had a bad case of sinusitis but became alarmed when his teeth began falling out.

A radiologist in New York City, Joseph Steiner, looked at Byers's radiographs and noticed some similarities between the developing bony lesions in Byers's mandible and those described in the deceased radium dial painters. Frederick B. Flinn, the prominent radium expert from the department of industrial medicine at Columbia University, was called in as a consultant and confirmed Steiner's suspicions: Byers's body was slowly decomposing, the result of massive radium intoxication from the Radithor. Flinn's conclusions were not made public, however, in part because other experts—including Byers's personal physician—refused to accept them.

By September 1931 the commission's investigation was well under way, and the ailing Byers was called to testify. He was too ill to travel, so a special attorney, Robert H. Winn, was sent to Byers's Long Island mansion to take the deposition. He later described the scene:

"A more gruesome experience in a more gorgeous setting would be hard to imagine. We went to Southampton where Byers had a magnificent home. There we discovered him in a condition which beggars description. Young in years and mentally alert, he could hardly speak. His head was swathed in bandages. He had undergone two successive jaw operations and his whole upper jaw, excepting two front teeth, and most of his lower jaw had been removed. All the remaining bone tissue of his body was slowly disintegrating, and holes were actually forming in his skull."

On December 19, 1931, the commission issued a cease-and-desist order enjoining the Bailey Radium Laboratories from continuing to market Radithor. The ruling came too late to do Byers any good. With his death in 1932, the commission reopened its investigation, and the FDA began campaigning for more sweeping powers. Medical societies took the opportunity to denounce all patent medicine sales, and calls for radium-control laws were voiced throughout America and Europe. The forerunners of the current regulations restricting the sales of radiopharmaceuticals to authorized users actually date back to the Byers affair. With the institution of the regulations, the radioactive patent



INVISIBLE RAYS from Byers's teeth left these eerie images on photographic film during the forensic examination that followed his gruesome death. Decades later a reexamination of the teeth and bones of Byers, Bailey and other consumers of Radithor showed significant levels of residual radioactivity. Although many individuals had quickly succumbed, some—including Bailey—managed to reach a ripe age, often without suffering any ill effects that could be attributed clearly to the radiation. This extraordinary variation in the effects of long-term radiation remains unexplained.

medicine industry collapsed overnight.

Bailey was never prosecuted for Byers's death, and he maintained that it was all a case of misdiagnosis. "I have drunk more radium water than any man alive, and I have never suffered any ill effects," he claimed. The Federal Trade Commission officially shut down his Radithor operation in late 1932, but by then the Great Depression had already taken its toll on Radithor sales, and Bailey had moved on to several new radioactive scams: the Bioray, a radioactive paperweight advertised as a "miniature sun"; the Adrenoray, a radioactive belt clip; and the Thoronator, a refillable radioactive "health spring for every home or office."

Hounded by the press and Newark public health officials, Bailey left town at the height of the Byers scandal and was finally located by reporters living at the nearby office-front headquarters of his Adrenoray Company in East Orange. When asked what kind of business was being conducted there, he said it was "a special advertising business" and that the name on the door "didn't mean a thing."

Bailey dropped from the limelight. He became the editor of the *Bloomfield Times* in New Jersey and wrote books on world politics and health. During World War II, he was an aircraft observer under the 1st Fighter Command. He also invented a method of swimming instruction for soldiers, devices for submarine detection and a mechanism for gunsight calibration. He may also have been a wartime manager of a small electronics division of IBM. He died on May 16, 1949, in Tyngsborough, Mass., at the age of 64. Although he died of bladder cancer, he never agreed that small doses of radioactivity were harmful, and he asserted that his health and spirits were excellent almost to the end. "I have developed," he wrote, "a philosophy of life that does not permit antipathies of any kind." Tax records from Tyngsborough list the total value of the Bailey family real estate as $4,175 at the time of his demise.

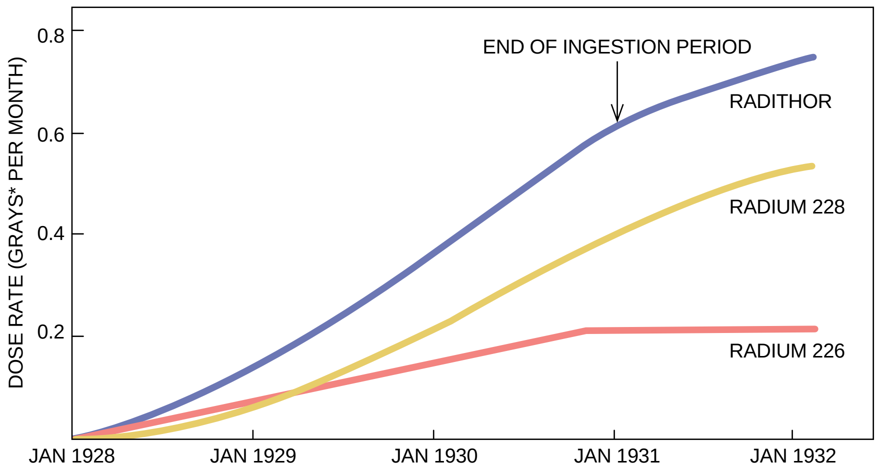
There is a curious postscript to this story. Soon after my library research on Bailey began, I realized that I was treading in the footsteps of a giant. Beginning in the 1930s, Professor Robley D. Evans, emeritus director of the Radioactivity Center at the Massachusetts Institute of Technology, had investigated a number of Radithor poisonings as part of his large study of the illnesses of radium dial painters and chemists. Evans and his team from M.I.T. and Argonne National Laboratory studied 29 patients (21 living and eight postmortem) and found wide variation in the distribution of radioactivity accreted in the bones and in the resulting consequences for health. Moreover, the toxic effects of these radiation loads appeared to vary greatly from one victim to another.

The data suggest that some individuals can tolerate high amounts of bone-seeking radioisotopes, such as radium. Perhaps these people are very efficient in excreting the substances or in producing protective hormones (such as granulocyte colony–stimulating factor and the interleukins) that stimulate the growth of blood cells when the body is exposed to radiation. Among these comparatively lucky consumers of Radithor was Bailey himself.

Among the materials that Evans and his colleagues studied in the 1960s were the disinterred remains of two Radithor drinkers whose clinical stories correspond so closely to those of Eben Byers and William Bailey as to constitute a positive identification. The bones of both men showed severe radiation changes and were still dangerously radioactive, almost 50 years after the individuals had drained their last bottle of Bailey's elixir. William Bailey, the dean of the radioactive quacks, thus played a major role in the beginning, the climax and the end of the era of mild radium therapy.

The study of the long-term effects of alpha radiation continues. Since my original publication of this research, other apparent victims of medicinal radium poisoning have contacted me, and my colleagues and I are studying their medical records and attempting to understand the radiation effects through experiments on rodents. Although the fad of radioactive patent medicines has long passed, some of the data it has bequeathed will continue to suggest answers to the problems posed by radon in the home and by nuclear waste in the environment.

These insights may also help workers to judge the costs and benefits of therapeutic alpha radiation, now being developed for cancer therapy. For, as Bailey noted with perhaps unwitting prescience and certainly with irony: "Radioactivity is one of the most remarkable agents in medical science. The discoveries relating to its action in the body have been so far-reaching that it is impossible to prophesy future development. It is *perpetual sunshine*."



** A gray is an absorbed dose of one joule per kilogram.*

SKELETAL RADIOACTIVITY in Byers is shown in graphs based on estimates of the potency of Radithor and the quantity consumed. Because some decay products are more radioactive than the radium from which they derive, the dose rate increased even after Byers stopped taking the medicine.

FURTHER READING

THE LATE EFFECTS OF INTERNALLY DEPOSITED RADIOACTIVE MATERIALS IN MAN. Joseph Aub, Robley Evans, L. Hempelmann and N. Martland in *Medicine,* Vol. 31, No. 3, pages 221–329; 1952.

RADIUM AND MESOTHORIUM POISONING AND DOSIMETRY AND INSTRUMENTATION TECHNIQUES IN APPLIED RADIOACTIVITY. Robley D. Evans in *Annual Progress Reports 1962–1968,* prepared at the MIT Radioactivity Center under contract AT(30-1)-952. National Technical Information Service, 1968.

NUCLEAR FEAR: A HISTORY OF IMAGES. Spencer R. Weart. Harvard University Press, 1988.

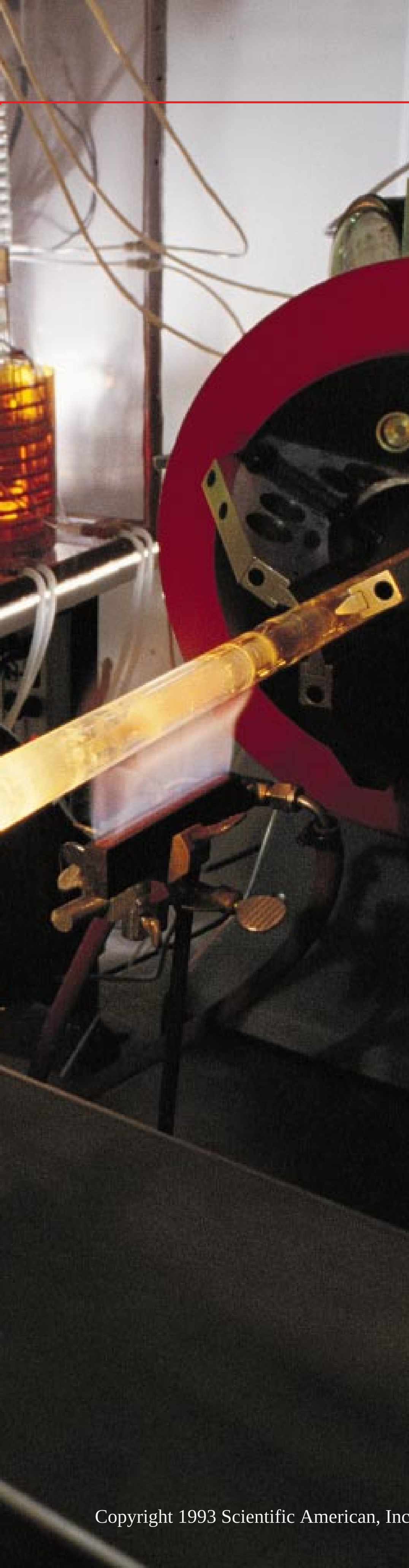
RADIOACTIVITY AND HEALTH: A HISTORY. J. Newell Stannard. National Technical Information Service, 1988.

RADITHOR AND THE ERA OF MILD RADIUM THERAPY. Roger M. Macklis in *Journal of the American Medical Association,* Vol. 264, No. 5, pages 614–618; August 1, 1990.

Trends in Communications

DOMESTICATING CYBERSPACE

by Gary Stix, *staff writer*



A technophile vice president and the information, entertainment and communications industries have discovered the source of next-generation electronic products—it's the network.

To prepare for a congressional hearing, Vinton G. Cerf sent out an open message on the web of computer networks called the Internet. Cerf, who is president of the Internet Society, asked users how they found the network to be helpful. Within a matter of hours, thousands of replies arrived in his electronic mailbox. "Without the infrastructure of the Internet, asking people about the network would not have been worth it, since the answers would have taken far too long to receive, and it would have been impossible to sort through the masses of paper," Cerf said.

Among the responses was a note from a professor at the University of Southwestern Louisiana recounting how his offer to teach a class using electronic mail relayed over "The Net" elicited responses from some 15,000 prospective students. A blind student told of tracking down archives of William Shakespeare's works that could be read through his speech synthesizer. Another story mentioned how President Bill Clinton's visit to a Silicon Valley computer manufacturer was quickly rendered into a stream of digital bits and transmitted over the network to hundreds of sites.

The creative ferment of the Internet—with its thousands of computer networks, comprising 1.7 million computers in more than 125 countries—is frequently invoked by the legislative legions in Washington who want to extend some version of electronic networking to every home, school, library and hospital in the country under the rubric of a National Information Infrastructure. Not surprisingly, the Internet's 20 percent average monthly traffic increases on its most heavily used segment have also tickled the fancy of media and communications megaliths entranced by an information service with growth rates measured on logarithmic scales.

The explosion of activity on the Internet has coincided with the ascendancy of a vice president who readily poses for photographs beside an ion-beam implanter as well as an unusual confluence of interests among computer, communications and entertainment giants. The migration to the Internet by universities, government agencies, community organizations and even business electronic mail users is seen as stirrings of mass appeal for electronic networking beyond the automated teller machine. "I see seeds of a National Information Infrastructure in Internet," says Robert W. Lucky, vice president of applied research at Bell Communications Research (Bellcore).

Moreover, the technical accoutrements are almost in place to extend computing technology into the television on the living-

GUIDING LIGHT for gigabit networks will be photonic technologies that will become the foundations for a National Information Infrastructure. An AT&T Bell Laboratories researcher monitors the fabrication of a fiber for an optical amplifier that is made by depositing a vapor of erbium into a glass rod.

room credenza—the same trend that saw it move from the mainframe data center during the 1980s to the office worker's desktop. Like the interstate highway system, the main arteries of the "information superhighway" envisioned by Vice President Al Gore have been under construction for the past decade. The flow on AT&T's transcontinental network tops out at 3.4 gigabits per second, or, in the popularized measurement, more than three encyclopedias' worth of text a second. Although data compression technologies can send a digitized movie over existing copper wires, lightwaves may prevail. Corning, the fiber manufacturer, estimates that roughly 2.3 million miles of fiber may be strung in 1993 by leading cable and local telephone companies, a 25 percent rise over last year. (From 12 to in excess of 200 fibers may fit into a single cable, so the actual number of miles of cable laid will be about 55,000.)

What telephone and cable companies propose to build, either together or as competitors, are the "on and off ramps," and the toll booths, to these electronic transcontinental thoroughfares. They will own, if Washington so allows, both the medium and its message units. For the consumer, the cost of admission will be not just the quarter for a pay call but a few dollars more for the video conference, the pizza order and every episode ever recorded of "My Mother, the Car." The $20 or $30 a month for a cable bill could double or triple, once "on-line connect" charges are added.

To garner the requisite expertise and billions of dollars in financing, the divested Baby Bells and the cable television companies have sheathed their swords to form multimedia *kereitsu* that make an art of the joint venture and strategic partnership. U.S. West teamed up with Tele-Communications, Inc. (TCI), the largest U.S. cable deliverer. Later the same regional telephone company announced an agreement with cable giant Time Warner. Another, Bell Atlantic, has set itself up to be a supplier of video pipes for local cable companies. AT&T is preparing to launch an interactive television trial with the cable company Viacom. And Southwestern Bell has purchased two cable companies near Washington, D.C. "The Mongol hordes are raging, but there is no Genghis Khan nor a David Sarnoff," observes Michael Hawley, a professor of computer science and electrical engineering at the Massachusetts Institute of Technology's Media Laboratory.

Boardwalk and Park Place

Indeed, futurists such as George Gilder of the Discovery Institute, a conservative West Coast think tank, assert that the inexorable pace of technological advance will encourage a wealth of competitors—direct-broadcast satellites, cellular television or hand-held personal communications devices—that will nibble away at the grand old monopolies held by the offspring of Ma Bell. In Gilder's view, the existing regulatory framework is antiquated and should be scrapped so that companies can form partnerships or engage in a free-for-all as they choose. But that has not yet happened. The cable industry has just been reregulated because of service problems, rate gouging and the failure of competition to emerge in local markets. And the basic connections to the home on the telephone company's local network are still a protected preserve.

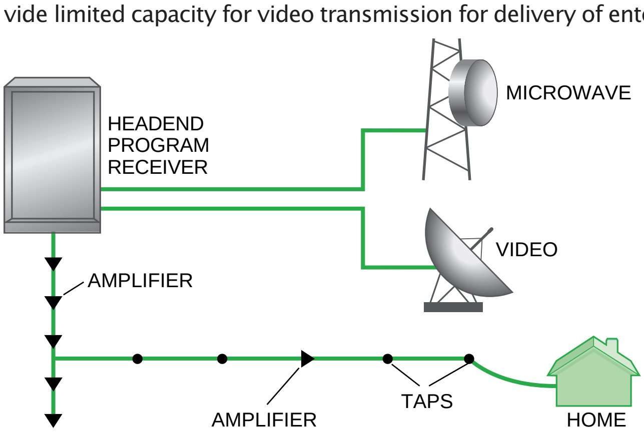
Even before Gore put forward the notion of an electronic superhighway, cable companies were beginning to use optical fiber to improve their existing networks. Their goal was to eliminate the string of failure-prone electronic amplifiers that are needed to boost the signal every 2,000 feet—one reason that cable systems often have outages. "In a large, urban cable television network, the distance to the furthest subscriber is about 40 amplifiers," says John Ryan, a South San Francisco communications consultant. "That's a maintenance nightmare." Just as important is the fact that the amplifiers induce noise into the network, which affects picture quality. This equipment also inhibits the two-way communication needed for ordering videos or products hawked on a home-shopping show.

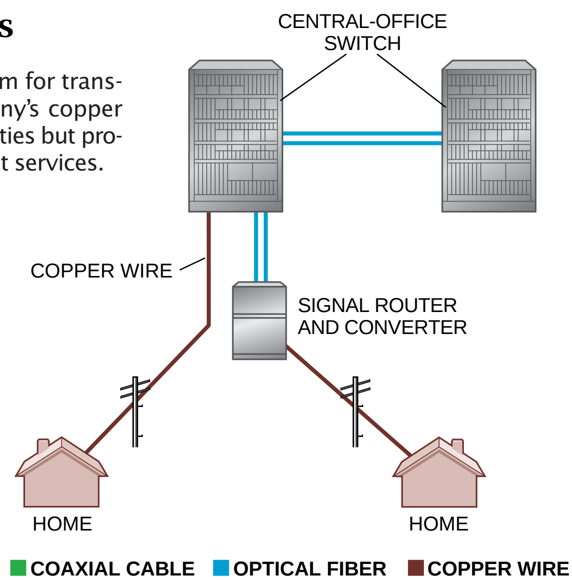
In the upgraded networks, optical cables reach a point a mile or less from a group of homes. Then existing coaxial cable provides the rest of the link. In raw communications capacity, this cable supplies the industry with an advantage over the narrow bandwidth in the telephone network's copper wiring. The coaxial cable can carry up to 1,500 channels of compressed video signals for short distances. The cables that constitute the physical pathways of this terrestrial television network already pass about 95 percent of U.S. homes and serve almost 60 percent of all households, according to market research estimates.

After the election of Clinton and Gore, some of these capital upgrade programs were politically corrected to become information superhighways. The "Infostructure Network" is the trademark that TCI has applied for. The company announced in April a four-year, $2-billion capital-spending plan to lay fiber

Today's Cable and Telephone Networks

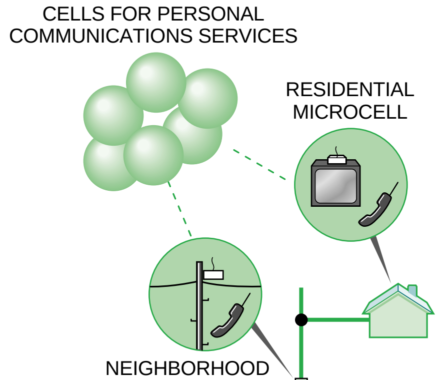
Cable television supplies a one-way, high-capacity broadcast medium for transporting videos to the home (*left*), whereas the telephone company's copper wires (*right*) furnish switched communications between two or more parties but provide limited capacity for video transmission for delivery of entertainment services.

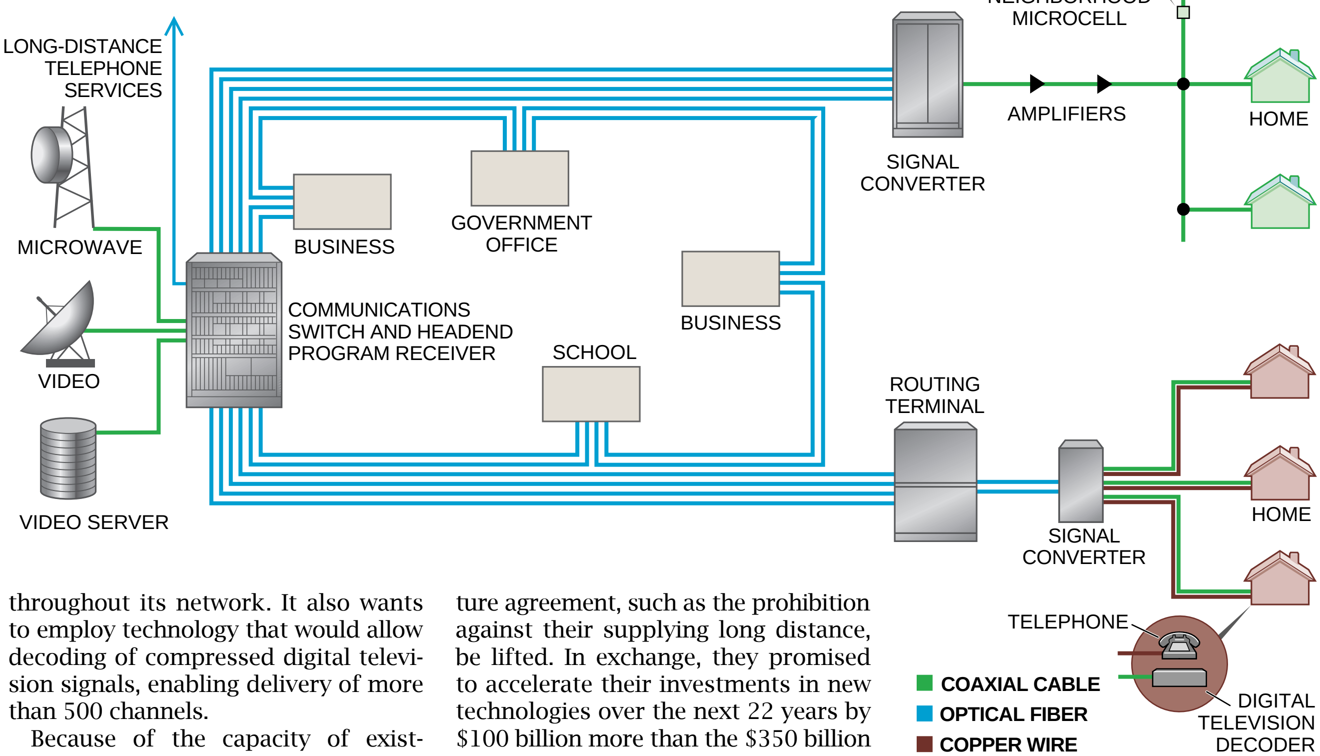




Evolution of Local Communications

A ring around the city will extend to businesses, schools and government offices as telephone and cable companies use optical fiber to combine voice, video and data for building interactive communications networks. Videos received by microwave or satellite, or stored in a local video server, will be channeled through a high-speed switch along with voice and data traffic. Telephone and cable companies will string fibers near a home, where they will be converted to electrical impulses. Telephone companies (*bottom right*) may complete the link with copper wiring for telephone calls and coaxial cable for television. Cable providers (*top right*) may use a coaxial connection for video and personal communications services (wireless networks) for telephone calls, although telephone companies are also investigating networks built on this cellular technology.





throughout its network. It also wants to employ technology that would allow decoding of compressed digital television signals, enabling delivery of more than 500 channels.

Because of the capacity of existing coaxial links to the home, the cable industry may be able to rebuild its network with fiber for $200 to $300 per customer, whereas telephone companies may need to spend $1,500, according to market and academic researchers. That differential may narrow, though, when cable companies must go beyond merely broadcasting their programming, as they do now, to furnishing switching services that can route a voice call or a video conference from one point on their network to another, as does the telephone company. Cable companies may also have to learn something about customer service. They have not been compelled to maintain the same level of service reliability as the telephone companies. Having access to 911 is deemed more important than being able to view "The Simpsons."

Meanwhile the regional telephone companies have been spending a lot of time in Washington since the election. Top executives from the Baby Bells met with Gore in mid-April to ask that remaining restrictions of the Bell divestiture agreement, such as the prohibition against their supplying long distance, be lifted. In exchange, they promised to accelerate their investments in new technologies over the next 22 years by $100 billion more than the $350 billion they had originally expected to spend. Reaction to such proposals has been mixed. "If you let local telephone companies into long distance, that's all you're doing—putting Humpty-Dumpty back together again," MCI Communications chairman Bert C. Roberts, Jr., said in a speech.

So far the Clinton administration has given no official response, but the progeny of Ma Bell have already gotten much of what they wanted. The Federal Communications Commission opened the door last year for telephone companies to carry video programming over their networks. But a prohibition against owning or licensing programs in the companies' own serving areas stands. After the FCC decision, New Jersey Bell, which has plans to pave virtually all of its network with fiber by the year 2010, struck an agreement with two cable companies to provide this "video dial tone."

Despite wider regulatory latitude, the Baby Bells are not satisfied. Bell Atlantic, the parent of New Jersey Bell, wants to become a cable company and has set up a separate subsidiary called Bell Atlantic Video Services. Along with Chesapeake & Potomac Telephone Company of Virginia, another operating unit, it sued the federal government last December, charging that the programming prohibition violates the company's First Amendment rights. The newspaper, broadcast and cable industries have filed friend-of-court briefs in opposition. Two bills in Congress would also release the telephone companies from these programming restrictions.

At the same time, the telephone companies' technical ability to compete with the cable networks is slowly improving. In many cases, they already have fiber pipes connecting their switching offices. Fiber constitutes almost 5 percent of total cabling (the rest is copper) in the Baby Bells' network. In extending these links, fibers will not stretch all the way to a home but to a curbside attachment point, or optical network unit, a few

The U.S. Lead in the Information Age

TELEPHONES	U.S.	JAPAN	EUROPE
LINES PER 100 PEOPLE	48.9	42.2	42.2
CALLS PER PERSON PER MONTH	43.4	46.1	48.7
CELLULAR TELEPHONES PER 100 PEOPLE	2.6	1.2	1.2
TELEVISION			
HOUSEHOLDS WITH CABLE (PERCENT)	55.4	13.3	14.5
VCR-RELATED EXPENDITURES PER HOUSEHOLD PER YEAR IN DOLLARS	44.6	35.3	14.1
COMPUTERS			
PERSONAL COMPUTERS PER 100 PEOPLE	28.1	7.8	9.6
DATA-BASE PRODUCTION (PERCENT OF WORLD)	56.0	2.0	32.0

SOURCE: Consumer Federation of America

hundred feet away. From there, coaxial cable will bring the video signals into individual homes; copper wires will carry in the telephone signals. A North Carolina company, BroadBand Technologies, has introduced the electronics for a fiber transmission and switching system that forgoes the complex modulation and circuitry used to broadcast cable programming. Similar to the technology used in some computer local-area networks for offices, it may be able to offer interactive video to the home for less than $500, the company claims.

Yet the telephone companies have still to deliver on the promise of some investments already made. A network that could seamlessly switch digital voice, data and video—the essence of an electronic superhighway—has been talked up for a decade by the Baby Bells as part of a scheme known as Integrated Services Digital Network. ISDN for the home would supply two digital channels, each 64,000 bits per second (the amount of bandwidth required for an uncompressed digital voice signal), and an additional data channel of 16,000 bits per second. Besides giving the consumer two telephone lines, these digital channels would dwarf the capacity of the analog modems that now transmit digital signals along telephone lines at speeds up to 14,400 bits per second.

The digital switches and software for ISDN are in place in many telephone company switching offices, paid for out of basic telephone rates. Although ISDN may furnish most of the capabilities of a broadband-fiber network, the telephone companies have lagged in deploying the technology at affordable rates. Today five states have ISDN tariffs targeted to residential customers, and costs can be high. Telephone companies say the service will become more widely available now that a set of technical standards is in place that lets switching equipment from one manufacturer work with that of another in routing ISDN's digital signals.

Mark N. Cooper, director of research for the Consumer Federation of America, argues that the telephone company's focus on optical-fiber technology will cost billions and lead to overcapacity for already underutilized digital switches. Nor is it necessary to build these broadband networks to keep the U.S. ahead of foreign competitors. According to the federation, the U.S. has the world's most advanced telecommunications infrastructure, outpacing Europe and Japan in the penetration of telephone lines, cellular telephones, cable television and personal computers. Especially significant is the fact that the U.S. population has absorbed many times the number of personal computers as have the Europeans or the Japanese. "The argument that pulling fiber will advance them ahead of us when nobody there has PCs is a non sequitur," Cooper says.

Low-Fiber Diets

The sense of urgency to deploy fiber may be further mitigated by asymmetric digital subscriber line (ADSL), a technique developed by Bellcore, the research arm of the local telephone companies. ADSL uses specially designed transmitters and receivers to reduce the interference and noise normally encountered on a copper line. It would permit transmission of compressed video signals over ordinary copper wire—at best, superior in quality to a conventional television picture. The ADSL transceiver can be attached to a single subscriber line without having to string optical fibers to an entire neighborhood.

The Electronic Frontier Foundation (EFF), an organization established by software entrepreneur Mitch Kapor, has become an advocate for these digital technologies because they allow a mail message or a digital video file to be routed from any sender to any receiver with the telephone network's switching capabilities—an inherently democratic communications medium. "It makes people into information publishers as well as consumers," says Daniel J. Weitzner, the EFF's senior staff counsel.

Such a switched architecture contrasts with the cable network that broadcasts one program to many receivers. Vestiges of this hierarchical system may persist even after cable networks permit two-way communication over their television networks or after the telephone companies build optical links, Weitzner says. "There's a concentration of enormous economic power behind this media distribution," he notes. "What we may be looking at is one-way entertainment coming down through the pipe with a little upstream capacity to select a movie or an item out of the Victoria's Secret catalogue."

Some networking experts believe that anything but broadband pipes will quickly be exhausted. "You only get a chance to build new infrastructure every once in a while," says Rick Weingarten, executive director of the Computing Research Association, which represents academic and corporate researchers. But one key legislator who agrees with the EFF's position is Representative Edward J. Markey of Massachusetts, chairman of the House Subcommittee on Telecommunications and Finance. He has stated that he would like to see digital services widely deployed before his retirement in 2015, the date that telephone companies would finish laying fiber throughout the local telephone networks under an accelerated schedule. "The goal is not fiber to the home but information to the home," Markey says. Before the year's end he hopes to introduce legislation that will require the regional telephone companies to file tariffs at affordable rates for digital services, such as ISDN.

Whether or not the National Information Infrastructure ends up being born optical or relies on digital technologies that can utilize existing copper wires and coaxial cable begs the question of who will pay—and how. And here, too, the ad hoc Internet has been drawn into the center of the debate over the appropriate role that government and industry should play in building and operating communications networks.

This past March executives from the major local and long-distance telecommunications companies signed a policy statement that outlined their vision of the place for government and the private sector in constructing networks. Arguing that government subsidies for building networks will deter commercial providers from entering these markets, the telecommunications industry broke down networking into two categories: "experimental" and "production."

Experimental gigabit networks could receive direct government funding. Production networks, those tapped into by researchers who do not require such advanced communications capabilities, would be built and operated entirely by the private sector. Federal funding for building the academic and research networks used for routine communications such as electronic mail would be phased out, although the government could subsidize users who purchase connections to the network from commercial providers.

Some of these suggestions are in line with what the government was already planning. In coming years, the National Science Foundation (NSF) wants traffic traveling over its "backbone" network—today the main conduit for the Internet—to be carried by commercial suppliers. Instead of the existing backbone, the agency has commissioned a high-speed network that will be limited largely to research on communications and supercomputers.

The telephone industry's well-defined categorizations have riled leaders of organizations that establish and manage the various "mid-level" networks run by universities, research institutes and libraries. The NSF, though, will continue for a time to provide a needed measure of funding for building and operating some of these networks. This dispute, it is feared, could ultimately disrupt the spontaneous, grass-roots quality that has enabled the Internet to grow into a worldwide network serving some 10 million users.

The telephone industry has been lobbying hard in Congress to make sure the public-private distinctions are retained in various bills to fund applications for health, education and library services under the rubric of the National Research and Education Network (NREN). If the Internet—or some outgrowth of it—evolves into a network that connects schools, hospitals and other public institutions, the telephone companies believe they may have a role in providing the kind of services they fur-

Stories from "The Net"

Earlier this year the Federation of American Research Networks, a nonprofit group that promotes use of computer networks in research and education, asked how the Internet can be of value to schools, health care and various public institutions. Here's a sampling of what people said.

Bruce Daley, teacher, Elaine Wynn Elementary School, Las Vegas, Nev.: "Paul Smith, a 24-year-old doctoral student from Latrobe University in Melbourne, Australia, was recently assigned to Casey Research Station on the continent of Antarctica. While stationed at the research facility, Smith used electronic mail over the Internet to describe his experiences of living and working in Antarctica to my third-grade class."

Norman F. Herzog, manager, engineering department, UNO Television, University of Nebraska at Omaha: "My daughter was scheduled for surgery in October of 1991 for correction of scoliosis (curvature of the spine). In late summer of that year, I decided it was important to learn more about scoliosis. A library catalogue search over the Internet led me to discover that another daughter had symptoms that could mean our family was affected by a serious hereditary disorder known as Marfan syndrome. I used a specialized Internet service, called WAIS, that let me search multiple data bases. The bibliographies led me to physicians who knew how to diagnose and treat it. The Internet may have saved my daughter's life!"

Allan G. Farman, professor, radiology and imaging sciences division, University of Louisville: "In dentistry, approximately 20 percent of all procedures require prior approval based upon the submission of radiographs. The Internet has proven very efficient and reliable for transmission of the relatively small amount of data in dental images. Transmission rates of 20 seconds have been possible for images between Louisville and Seattle, Vancouver, Torino, Italy, and Amsterdam, the Netherlands. Image analysis has shown that there is no loss of information during transmission."

Ferdi Serim, computer teacher, West Windsor/Plainsboro Regional Schools, Plainsboro, N.J.: "I strive to make the classroom a place where children feel comfortable exploring, and through the Internet our "playground" is truly global. One example was the Chernobyl exercise. It lasted only 36 hours, but it was profound. On Monday, at noon, a message came over the Internet stating that 40 students from the contaminated zone in Chernobyl would be visiting a health spa in England. On Tuesday morning, my first two classes used our word-processing skills to craft "get well" messages for the children. It took a little more than an hour to write 50 messages and 30 seconds to send them to England. The next morning we received a reply from the Russian teachers. They were amazed that U.S. children knew and cared about the plight of their students. History, science and language arts blended to make a personal sense of current events tangible to our students in a way otherwise impossible."

Ferdi Serim's computer class

nish on their large voice networks: reliable access and sophisticated billing capabilities. To make their point, they have named managers with titles like marketing director for Internet/NREN and executive director for National Information Infrastructure Initiatives. "This is a delicate thing," says Bellcore's Lucky. "The Internet is a wonderful sociological experiment. The question is, can it be scaled up to an industrial-strength network and still retain its flavor?"

Internet veterans point out that telephone companies have yet to deliver. The Baby Bells generally do not supply the types of expertise and services—software to route messages on the network, for example—that users expect from smaller networking organizations. "These markets may not have a lot of money in them," says Weingarten of the Computing Research Association. "They're not well organized, and they are not attractive for the telcos to go hitting on." Others guess that the communications services market may, in fact, exceed $100 million nationwide, but the telephone companies' staidness is keeping them away. "They don't have the trained staff—they don't have the experience in operating these kinds of networks," Cerf charges. "What they're offering is a pack of excuses instead of getting in and getting their feet wet."

While sparring continues over the Internet's future, the media-marketing empires are forging billion-dollar deals. Their reason is not solely an altruistic urge to better serve needy schools and hospitals but the belief that traditional broadcast and print media will merge into a multimedia cloud that can float into the consumer's head during the more than seven hours a day the television is switched on. "The country needs an information supermarket, not an information superhighway," says Arno A. Penzias, who is AT&T Bell Laboratories's vice president for research.

Yet no one knows whether consumers are truly interested in foraging through a video wilderness of 500 to 1,000 channels. The idea that people want a virtually infinite choice of programming, not 10 or so channels, strikes some traditional broadcasters as an absurdity. "They won't buy the accountancy channel, the plumbing channel or the hardy perennials channel," Howard Stringer, president of CBS/Broadcast Group, told an audience at a broadcasters' convention this past April.

Reasonable people will differ. "There's a magazine called *Guns and Ammo,*" Penzias says. "There are people who will sit there all day reading about a bullet. As far as they're concerned, if there is a channel devoted to guns, their programming will improve."

This kind of frenzied activity has a cyclical quality. In the late 1970s the Knight-Ridder company assembled in great secrecy a team of editors and technologists to create Viewtron, a service with community news, banking services, travel reservations—many of the things that interactive television promises, with the addition of the video. A fatal mistake then was giving customers a terminal from AT&T that was incompatible with the Apple and IBM computers that were gaining widespread acceptance.

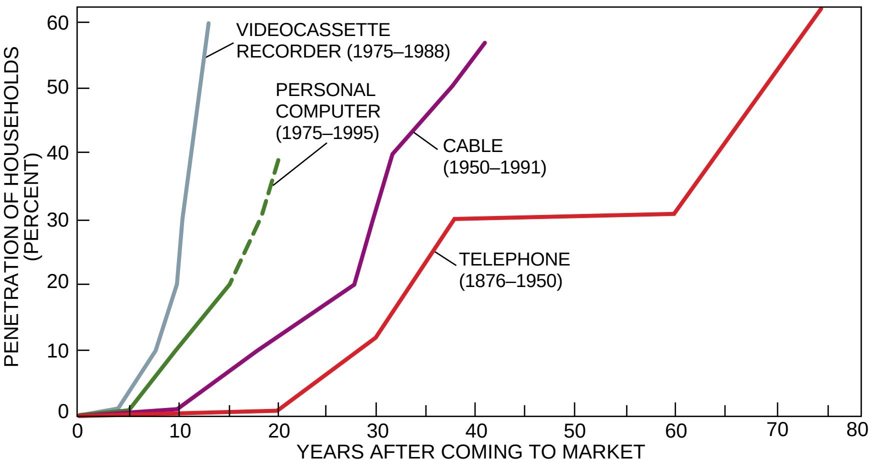
In 1986, after spending $50 million, Knight-Ridder dropped its experiment. Viewtron subscribers primarily logged onto the system for electronic mail and bulletin boards that now draw personal computer aficionados to on-line services like CompuServe and Prodigy as well as to the Internet.

Déjà Viewtron

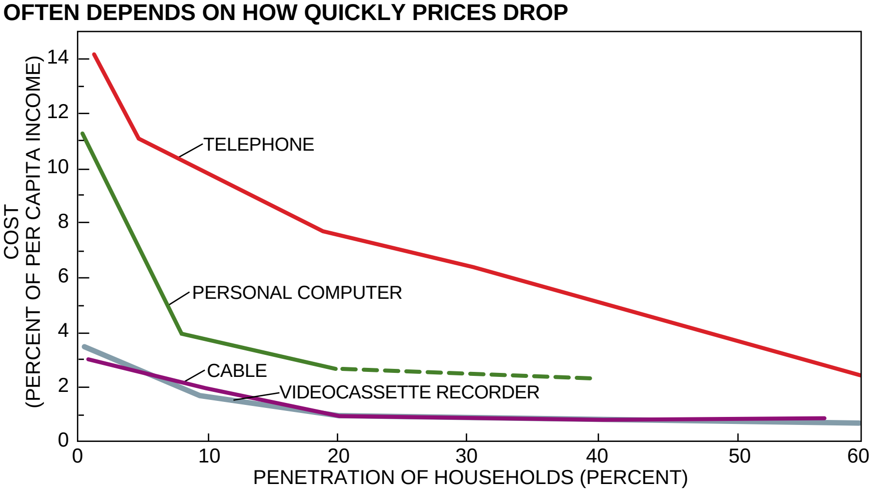
But spurred by improvements in the cost and performance of computing technology, Knight-Ridder has come back for more, although it is treading cautiously. For some time, it has supplied on-line text versions of its newspapers. Last fall it set up an information design laboratory in Boulder, Colo., that is exploring the concept of an all-electronic newspaper that goes beyond sending plain text files by modem. This time the group has no plans to commission special hardware, as it did before from AT&T. But it is experimenting with early versions of a two-pound, tabletlike computer consisting of a flat-panel screen about the size of a standard sheet of typing paper.

A penlike pointing device lets a reader unpack a full news story from a page full of abstracts, similar to the *Wall Street Journal's* front-page news summaries. These stories could be beamed to an electronic tablet over a radio link. The reader might then access a number of other stories, old or new, in the newspaper's data base: sidebars on dif-

HOW FAST THE INFORMATION AGE COMES HOME...

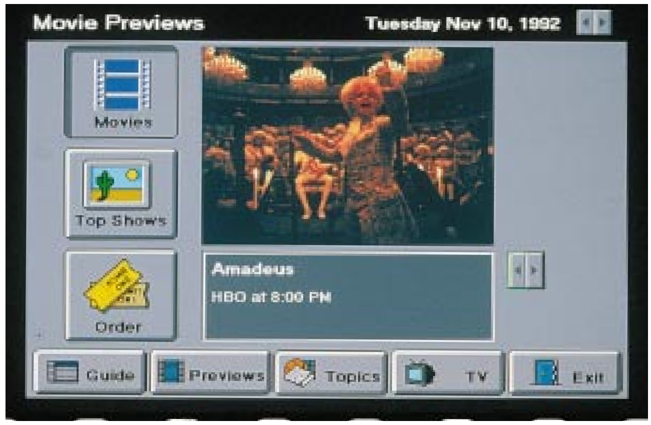


OFTEN DEPENDS ON HOW QUICKLY PRICES DROP



SOURCE: Consumer Federation of America

ADOPTION OF A TECHNOLOGY by most U.S. homes can take little more than a decade or as long as 75 years (*top chart*). Diffusion often accelerates as costs drop (*bottom chart*)—a reason interactive services may now prove more successful.



TALKING BACK TO THE TV may become the great American pastime if electronic video stores and newspapers take hold. Roger F. Fidler of Knight-Ridder's Information Design Laboratory is leading an effort to create a prototype electronic newspaper using a computerized tablet (*left*), and Microsoft has developed a working model of an interactive movie guide that would run on a cable television version of its popular Windows operating system (*right*).

ferent ethnic groups in Bosnia, maps of the region and short video clips.

The secrecy that surrounded Viewtron is gone. Today there is a premium on risk sharing. Knight-Ridder teamed up this past spring with several other media companies—Gannett, Hearst and Times Mirror, among them—to support the "News in the Future" program begun by the M.I.T. Media Laboratory with $6.5 million over five years. "We're trying to make news understandable to the machine," says Nicholas P. Negroponte, the Media Lab's director.

But the integration of artificial-intelligence techniques into the process of editing and reading a newspaper is poorly understood and controversial. The Media Lab has been a leader in the development of "agents," a kind of software version of the television character Max Headroom. The agent, in principle, will watch a reader, note his or her preferences and then go out and dredge up news items from the terabytes of information in data banks, the makings of a personalized newspaper. "The idea of an intelligent agent asking whether you would like french fries with that burger is something that people don't really like," says Diana Hawkins, a former Media Lab researcher, who heads her own firm, Interactive Associates.

To Roger F. Fidler, a member of the original Viewtron team, agents may yield an improved method for reporters and editors to ferret out information from large stores of information. The actual task of editing a newspaper—the decisions by the editor about what news is important—is a form of nonsequential organization that surpasses what may be accomplished by an agent or other computer techniques to link data. "One of the things not well understood by technologists is that editors spend a good deal of time anticipating what readers want," Fidler says. "People think you can just hit this infinite 'more' button, and things will just materialize out of cyberspace."

Overlaying proved and familiar elements—lead story, headline and lead paragraph—onto the amorphous structure of the electronic medium may turn out to be the best way to wean readers away from cellulose, Fidler believes. The saving grace of the electronic newspaper may be institutional survival.

During the 1960s and 1970s, newspapers decided to invest in computer technology to reduce the high costs that were making them uncompetitive against radio and television. "Today newspapers up to the printing press are digital," Fidler says. "The vestige of the industrial age is the printing press, and that is where the next opportunity lies to make enormous savings."

Despite the rush toward digital media, Penzias's vision of an information supermarket may be years away. And it may bear little resemblance to an at-home video store or an electronic news broadsheet. But what has begun to emerge are the bits and pieces of tools for handling any type of digital information, particularly the intensive processing and storage needs of video and film. These trusses and pillars of the information infrastructure could help spur a reemergence of a U.S. consumer electronics industry, since most of the technology is based on the innards of personal computers and workstations.

Moreover, the information revolution will almost certainly be televised. Simply concentrating on building TV sets that produce a crisp, photograph-quality image, once thought to be the cornerstone of the next generation in consumer electronics, is already beginning to take a backseat to incorporating computing power into home video displays. A laggard in getting involved with high-definition television (HDTV), the U.S. came up with digital technologies that ultimately won out over analog transmission systems being considered as standards in Europe and Japan. In 1990 the General Instrument Corporation devised an ingenious HDTV prototype that combined a variety of compression technologies to cram a digital video signal into the six megahertz of bandwidth allocated to broadcasters.

High-Definition Computing

Other companies and institutes subsequently turned their efforts to developing digital systems. At the same time, the computer industry argued for standards that would make the new sets compatible, in all respects, with the computer. This should encompass the way the cathode-ray gun paints the lines on the screen, the shape of the picture elements and the encapsulation of data into small packets. "The next generation of technology should take into account up front that the television is a digital display device," says Lee McKnight, who directs M.I.T.'s program on digital open high-resolution systems.

At the National Association of Broadcasters meeting in mid-April, Apple chief executive John Sculley suggested that HDTV's bandwidth could be broken up into smaller channels as needed, an "intelligent data stream" for programs with multiple camera views or text inserts. A viewer might look at a newscast on the environmental impact of a new dam by clicking a remote control button to choose the viewpoint of a construction worker or that of a fisherman. "I think that may be at least as interesting if not a more interesting opportunity during the 1990s than HDTV itself," Sculley noted.

By late May the computer companies seemed to have gotten at least some of what they sought. The major groups developing systems for testing under the FCC's aegis announced they had agreed to put forward a single digital standard for HDTV. Their proposal is a compromise that accommodates both interlaced scanning—one in which every other row is drawn on the screen by an electron beam—and the progressive scan (in which each row is successively traced), the format desired for compatibility with computers.

The interlaced display causes enough flicker to produce eyestrain for someone looking at text or graphics on a television screen. "Ultimately, your entertainment screen is going to have to be a display for all sorts of data access," Penzias declares. "Under those circumstances when you go to do some data processing, basically the display has got to be able to support a video that may be a dirty movie or data visualization, but it's got to be able to do both."

To do that, others are figuring out ways to imbue the household television with the ability to process the protoplasm of digital bits. What Gilder of the Discovery Institute calls a "telecomputer" is already provoking a competitive struggle. As with the personal computer, which undercut the primacy of central mainframe processing, the design and the amount of intelligence in the box on top of the television will determine how much control the consumer has over interactive television.

Dominant hardware, software and service suppliers want to make sure that their products become market standards. Microsoft, Intel and General Instrument are incorporating the equivalent of an IBM PC running a special version of the Microsoft Windows operating system into a cable converter box. In June, Microsoft was negotiating with TCI and Time Warner about joint development of an operating system for interactive television. Another company, Silicon Graphics, plans to make a television set-top box with the fast-processing chip used in its workstations.

The myriad deals being struck on all fronts are already roiling the regulatory waters. In March, Ameritech, a regional telephone company, petitioned the FCC to begin a process that could result in the lifting of legal prohibitions against offering long-distance service. In exchange, Ameritech would open that bastion of monopoly, the local telephone network in the five states it serves, to competition. Each part of the telephone network could be purchased separately: a customer could order a dial tone from another provider while still leasing the line itself from Ameritech and retaining the same telephone number. Columbia University telecommunications economist Eli M. Noam has suggested that "personal systems integrators" might be needed to make sense of the various options for consumers.

Ameritech and other telephone companies argue that they have nothing to lose: their businesses are under siege. There is already competition for yellow pages, operator and 800 service, pay telephones and regional toll calling. So-called competitive access providers can connect a large business directly to a long-distance carrier at an often lower rate than can the local telephone company. In Houston, for example, U.S. West could be competing against Southwestern Bell in helping Time Warner convert its cable networks to two-way communications systems.

This technological assault, predict analysts such as Noam, erodes the tradition of the common carrier—the obligation to maintain a public viaduct for a free flow of information, open to all comers. Competitors to monopoly common carriers such as the telephone companies can offer a lower pricing structure because they do not have to serve all customers on an equal basis. In tandem with the notion of common carriage, federal and state regulators have required that telephone companies furnish service in high-cost rural areas and offer affordable rates to low-income customers. In turn, urban and middle- to upper-income customers have paid more to ensure universal service.

Civil libertarians worry that having a common carrier like the telephone company own both the medium and the content—which, except for video programming, is now permitted—may stifle both competition and free speech. "Providers of a conduit that have an economic interest in the content running through that conduit are likely to favor themselves in unfair and anticompetitive ways," says Nicholas Johnson, an FCC commissioner during the Johnson administration. "It doesn't matter whether it's coal on a railroad, petroleum running through a pipeline or cable television on a telephone wire."

Consumer advocates argue that competition is more a specter than a reality: competitive access providers that connect businesses to long-distance carriers account for far less than half of 1 percent of the $95-billion local telephone market. Safeguards, they assert, are needed against monopoly predations. "It's like saying that technology will win over everything, so we should just back away," says Gene Kimmelman, legislative director of the Consumer Federation of America. "In 1984 we were told that cable should be deregulated because there were satellites, and all kinds of technology was coming, so we pushed regulation out of the way, and cable became an entrenched monopoly."

Congress, regulators and the courts have yet to grapple with the rewriting of communications laws needed to bring common carriage and universal service obligations to cable companies and other prospective suppliers of two-way communications. A multimedia information superhighway also suggests the need for a substantive rewriting of public broadcasting law. "We're really afraid that all we're going to be left with is a virtual shopping mall," says Jeff Chester of the Center for Media Education, which advocates the taxing of the various new media to pay for noncommercial programming.

Information for All

The idea of universal service may eventually encompass information. Access to the network may be required—not just for calling the fire department or for ordering pizza but for voting and participation in town meetings. "Information stamps" might be parceled out like coupons for buying food. "The equity issue is a very important one in this whole scheme," says Charles M. Firestone, director of the Communications and Society Program for the Aspen Institute. "These technologies could bridge the gap between rich and poor, and they could also, I'm afraid, widen that gap."

The tug and pull of competing interests may mean that little gets done on Capitol Hill or in the state utility commissions during the current sessions. Congress is likely to pass legislation to permit the auctioning off of 200 megahertz of radio spectrum to the private sector for wireless technologies such as personal communications services. Bills for networking for schools, libraries and health care may get through. The Clinton administration was also preparing to appoint a cross-agency task force to deal with computing and networking.

But a definitive road map for a National Information Infrastructure is still lacking. In its absence, the vitality of a public medium like the Internet could be jeopardized as powerful media interests jockey for position. "Anything is a danger in cyberspace," Representative Markey comments. "There are no rules. It's the Wild West." And claim stakes on the electronic frontier may be sold in 30- and 60-second spots.

Body English

Controlling computers with twitch and glance

STEVE MUREZ *Black Star*

MOTION BECOMES MUSIC as sensors on composer Atau Tanaka's arms transmit muscle signals to a computer that controls electronic musical instruments.

Arm waving, daydreaming, jaw clinching, staring: computers provoke such behavior in the best of us. But until recently, computers themselves remained unmoved by our antics. Now a neurologist-turned-entrepreneur is building machines that let you make music with your biceps, move cursors with your eyes and blast your astrofighter into hyperspace by setting your mind out to pasture.

Hugh S. Lusted is fulfilling a boyhood fantasy through his infant company, BioControl Systems in Palo Alto, Calif. "Way back in high school, I thought, 'Wouldn't it be great if you could make a machine that could control things directly from your mind?' It's not a new idea—in science fiction." But the old idea has materialized into a handful of new products that connect computers to muscles, eyeballs and brains by sensing and deciphering the electrical signals these organs emit.

BioControl was born out of an interest in music shared by Lusted and Benjamin Knapp, an electrical engineer at San Jose State University. They met on a cochlear implant project in 1987, when Lusted was a neurophysiologist at the Stanford University School of Medicine. "We were making a bionic ear," Lusted recalls, "implanting electrodes into deaf volunteers to stimulate the auditory nerve directly."

After hours, the two scientists assembled the circuitry and software needed to control electronic musical instruments with information teased out of electromyograms, readings of the electrical discharges that muscle fibers produce when they contract. Within months they had a working prototype. With electrodes strapped to their arms, the researchers could play invisible violins by bowing the air. Flexing one arm raised or lowered the pitch; tension in the other arm changed the volume.

Encouraged, Lusted and Knapp turned their signal-processing techniques on the electro-oculogram (EOG) signal produced by the eyes, in the hope of creating a hands-free computer mouse. "The eyeball is a battery, a dipole," Lusted explains. "The retina is negative in relation to the cornea, because the retina is metabolically active. We can detect the movement of that dipole in an electric field"—and not just movement but also depth. Because the eyes converge as they focus, the EOG can be used for three-dimensional as well as two-dimensional control of computer cursors, perhaps even of robots. And "because the EOG signal gets larger when the light is brighter, we can make use of the light level to get information about what the user is looking at," Lusted adds.

Glances proved more furtive than flexes, however. "Historically, the EOG was considered unusable because the signal drifted," Lusted says. "But that was just an electrode and electronics problem," which BioControl claims to have solved with hydrogel sensors developed by New Jersey–based Alternative Design Systems. Lusted likens the texture of the reusable sensors to contact lenses: "They're mildly sticky but leave no residue." Velcro bands wrapped around a limb or forehead hold the electrodes in place.

With muscles and eyes under (or rather in) control, Knapp and Lusted lunged for the pinnacle: brain waves. The same headband sensors that read eye movement can also pick up an electroencephalogram, reading the waves of changing potential radiated by brain neurons as they fire. "Ideally, you could think the word 'on' and turn on the lights," Lusted muses. "We're a long way away from that kind of pattern recognition, but that's the direction we're headed."

BioMuse, released in October as the maiden and flagship product in BioControl's line, can indeed turn lights—or anything else—on and off in response to brain waves (though not to specific thoughts). "It detects the transition between alpha- and beta-wave activity," Lusted explains. "For most people, whenever you close your eyes, you go into a little alpha-wave epoch. With a bit of practice, you can produce short alpha-bursts just by thinking black."

Based on a digital signal–processing chip produced by Texas Instruments, BioMuse incorporates all of BioControl's algorithms in one box. The device can decode signals sent simultaneously by up to eight sensors—enough to wire your limbs, eyes and brain, with a channel leftover for voice recognition. Although Lusted sees myriad applications for the device, at $10,000 for a three-year license, most current customers use it as a research tool.

There have been exceptions. Atau Tanaka composed and premiered a 2-minute musical performance work for solo BioMuse last summer while he was a graduate student of music at Stanford.

"I'll Trade You a Wallaroo for an Aardvark . . . ."

What do Reggie Jackson, Jeffrey Dahmer and the Ledbeater cockatoo have in common? They have all been immortalized by the $2-billion-a-year trading card industry. In the name of preserving endangered species, the San Diego Zoo has just released a trading card series that could make echidnas and Tasmanian devils into household words, just as they have kept alive the names of baseball stars and mass murderers.

A portion of the proceeds will go to the zoo's Center for the Reproduction of Endangered Species (CRES), the multidisciplinary institute whose projects include a "20th-century ark" that maintains frozen genetic material from some 350 endangered species. The trading cards are part of an effort by the Zoological Society of San Diego to make the center's growing $2-million budget independent of gate receipts, which declined last year because of a lackluster economy that slowed tourism. "We want to take CRES off that roller coaster and let it stand on its own," says Jeff Jouett, a zoo spokesman.

But achieving autonomy will require extending the zoo's hard-sell marketing tactics beyond southern California. The cards, which are already being distributed nationwide at grocery stores—you can buy them at 7-Elevens in Oklahoma—were printed by CARDZ, the same company that brings you the Flintstones and the "Tales from the Crypt" trading series. On their foil wrappers the "World Famous San Diego Zoo Animals of the Wild" cards show either a gorilla or tigers against a circus-yellow background. (The zoo cards are printed on paper, half of which is recycled fiber and another 10 percent of which is "post-consumer waste.")

The zoo has been run more like a business than have other parks, for which zoo officials are unapologetic. The reasons for the theme park–like tone are clear: Disneyland, Universal Studios and Sea World, to name a few. Only 2 percent of the zoo budget comes from local property tax receipts, compared with half or more for some other municipal zoos. "The recurrent comparison is to Disneyland," Jouett says. "It's ironic because Walt Disney came to the zoo for information on how to start Disneyland."

The zoo's so-called team management approach to administration of the zoo and its companion Wild Animal Park has been written up as a case study for the Harvard Business School: a team is responsible for everything from animal feeding to information management for a particular bioclimate exhibit. And the zoo has recruited patrons from Los Angeles with racy local radio spots that have enjoined Angelenos to "come check out the naked mole rats." The Los Angeles Zoo looks with envy at the $3 million the San Diego Zoo spends on advertising, which is about a quarter of the more northerly park's entire annual budget.

The zoo's aggressiveness has drawn formal protests from the World Wildlife Fund for trying to procure a federal permit to bring in endangered pandas from China for breeding at the zoo. The presence of pandas—and accompanying sales of stuffed toys and other memorabilia—could draw millions of dollars in added revenues.

Humble trading cards are a tiny step toward economic independence for CRES. The zoo's deal with CARDZ to market the 110 cards, ranging from tuataras to wallaroos, will yield only $50,000 and a share of future royalties. But other resources of major zoos, such as the San Diego Zoo's 750,000-item photograph collection, may be viewed as potentially lucrative forms of intellectual property by legions of entertainment companies seeking photographs, video footage, text, graphics and other display materials for the emerging technology of multimedia.

Indeed, a well-established software company, Software Toolworks, used some of the zoo's photograph and video collection to produce a CD-ROM, or compact disc/read-only memory, that can be accessed by a computer. The disc is a narrated tour through 350 zoo exhibits.

The growing species preservation mission has caused other influential institutions to recast their image, with an eye toward broadening marketing appeal. The New York Zoological Society changed its name earlier this year to NYZS/Wildlife Conservation Society. A trading card series issued last year that included photographic contributions from the parent of the former Bronx Zoo (now the Bronx Zoo Wildlife Conservation Park) has been distributed at the likes of Toys-R-Us. Paying the zoo's geneticist may depend on how well an institution's marketing wizards can transmute the spotted cuscus into a cash cow. —*Gary Stix*

JASON GOLTZ

WILD CARDS are meant to grab a piece of the $2-billion-a-year trading card industry.

He titled the piece "Kagami," or "Mirror," because "the music is directly reflective of body elements," he says. Wearing two wireless armbands, Tanaka controlled via his choreography the timbre, pitch and sonic position of didgeridoo, bell and drum sounds.

More typical is Volvo, which licensed two BioMuse units for its automotive researchers. "They are monitoring drivers for fatigue and stress in their automobile simulators," Lusted says. "They are also interested in using the eye controller for heads-up displays in windshields. The idea is to keep your eyes up rather than looking at the instrument panel. It's a safety issue."

Realizing that the research market alone will not likely sustain the company long, BioControl is scaling down its technology to fit a few lucrative niches. Lusted claims he already has prototypes for four products, all derived from BioMuse. "We hope to be able to deliver them in at least small numbers by Christmas," he says.

The simplest, called BrainMan, is a brain wave–controlled switch for video games. "It will just be a toy," Lusted emphasizes."We're not making any great claims for it. But we are making a statement in the market: this technology is available, it's practical, and we can make it cost-effective." An eye-tracking video game controller (the "fire button" is a jaw clinch) and a muscle-controlled device for musical instruments are also in the works. All three products could cost less than $150 if mass-produced.

That is a big if. BioControl's overtures to video game leaders Sega and Nintendo have not been warmly received. "They're only interested in something that is linearly better than what they have now," Lusted complains. "These controllers are not faster or better; they're different."

Indeed, manufacturing and distribution partners appear to be a sine qua non for all of BioControl's designs. Funded by seven private investors, the company is hardly swimming in capital: it lacked a full-time president until last year, when Lusted left academia. "My research grant expired, and I could finally afford to pay myself a salary," he confesses.

For its fourth product, called Enabler, BioControl has the advantage of an enthusiastic ally. David J. Warner of Loma Linda University Medical Center has used BioMuse's muscle- and eye-tracking capabilities (which Enabler integrates in an add-on board for personal computers) to help diagnose and treat several disabled patients. "I'm in touch with a lot of cool stuff," he says. "But the BioMuse is one of the single most interesting interface technologies I've seen."

Enabler would allow people with even severe paralysis to operate a computer. Warner has tried the technology with three patients. Crystal, an 18-month-old quadriplegic, cannot even breathe on her own. But within minutes of donning a headband, the toddler learned how to move a smiley face around a computer screen with her eyes. Older quadriplegics could use an eye tracker in combination with existing word- and sentence-completion software to write.

Warner also hooked a paraplegic adolescent up to the BioMuse. After placing electrodes on the boy's dysfunctional arms, Warner says, "we gave him a rock-and-roll guitar sound, and he turned into Jimi Hendrix for 10 minutes—so much so that he broke into a sweat. This is somebody whom we couldn't get to do anything; the increase in motivation was profound."

Not so profound, perhaps, as the technology's impact on Dennis, a 38-year-old man whose heart attack deprived his brain of blood for 10 minutes. The extensive damage left Dennis, doctors thought, in a "persistent vegetative state, which basically means a body with a brain stem," Warner says.

On an instinct, he got permission to connect Dennis to a BioMuse. After locating a relatively quiet shoulder muscle, attaching electrodes and running baseline tests, "I gave him a specific verbal command to raise his shoulder. And we could clearly show that after a few seconds, there was a contraction orders of magnitude greater than any random noise." After several repetitions, Warner says, "I went to the neurologist and said, 'Look, somebody's here. I want his diagnosis changed.' "

"Dennis still has very low rehabilitation potential," Warner admits. "But the difference between a diagnosis of persistent vegetative state—which means nobody's home—and a diagnosis of severe cerebral hypoxia—which means that somebody's home but they're pretty messed up—is the difference between saying there is a soul there and there is not a soul there. Now when the nurses come in, they acknowledge that Dennis is aware. So if there is any potential for quality of life, we've increased that."

A stirring story alone will not win BioMuse approval from the Food and Drug Administration for regular use as a medical device, however. "Several people with reputations are going to have to stand up and do a study," Warner concedes. That takes funding; so far there is little. But in a changing health care environment, he points out, a technology "that could actually reduce costs by automating therapies" may yet garner support. At the very least, it promises some fascinating new toys. *—W. Wayt Gibbs*

Blood Money?

Critics question high pharmaceutical profits

When President Bill Clinton finally gets around to announcing his proposals on health care reform, measures aimed at limiting prices of prescription drugs will be only a small part of a large and complex package. But drug prices are the central issue for the pharmaceutical industry, which sees in the smoke signals from the White House the specter of controls that could slash revenues. The result, the industry warns, would be less research and development and, so, fewer new drugs.

A myth, replies Senator David Pryor of Arkansas, chairman of the Senate Special Committee on Aging and the sponsor of legislation that last year sought to control the price of prescription drugs. Pryor, whose Prescription Drug Cost Containment Act won the backing of then governor Clinton before it died with the last Congress, contends that "skyrocketing profit levels" in the drug industry must be checked. Over the past 12 years, Pryor asserts, drug prices have been increasing far faster than the rate of general inflation, even the 8 percent average rate of medical inflation. Whether the administration proposes mandatory controls or, as many observers think likely, a voluntary code, pressure on drug prices is sure to increase.

Pharmaceuticals accounted for about 10 percent, or some $70 billion, of the national health care bill in 1990. Manufacturers like to point out that drugs are good value when set against the alternatives that they can replace, such as surgery or prolonged illness. Although some blockbuster drugs do make many hundreds of millions of dollars, the Pharmaceutical Manufacturers Association (PMA) says the potential for high rewards is necessary to justify the risky and drawn-out process of developing original products. Research and development by the association's more than 100 research-based member companies will, it says, reach $12.6 billion this year, an estimated 16.7 percent of sales and exports. In any case, says David Emerick, a spokesman, the rate of drug price increases is now lower than it has been for 15 years and is roughly equal to growth of the consumer price index.

Yet in February a major study by the congressional Office of Technology Assessment (OTA) corroborated published academic research concluding that the industry is indeed a highly profitable one: between 1976 and 1987 returns in the pharmaceutical industry as a whole were

2 to 3 percent higher than in other industries. “Drug companies are very profitable by conventional accounting standards,” says Henry J. Aaron of the Brookings Institution. “If you treat research as a depreciating asset, the profits are more reasonable—but still pretty good.”

Pharmaceutical development is undeniably risky. The OTA calculated that the average after-tax cost of developing each drug that reached the market during the 1980s was $194 million (in 1990 dollars) and that the process took an average of 12 years. Each new drug introduced between 1981 and 1983 returned on average $36 million (after taxes) more than it took to develop, a surplus corresponding to just 4.3 percent of the price of each drug over its lifetime.

During the 1980s, revenues increased rapidly, so the industry stepped up research—by about 10 percent a year, according to the OTA. But even without price controls, the outlook for revenues is cloudy. Competition from manufacturers of generic drugs (who do little research but produce copies of brand-name drugs when their patents expire) is getting stiffer. “Patents worth $15 billion will expire over the next several years,” notes Frederick M. Scherer, a professor of economics at Harvard University, who chaired the OTA study’s advisory panel.

In addition, a 1984 law that reduced the amount of testing required for generics is starting to shorten the period during which drug developers have a captive market. Efforts to contain health care costs are also having an effect. “We’ll probably continue to see more health maintenance organizations,” Scherer predicts. “This will mean more pressure to prescribe generics.”

On the other side of the equation, costs of research are accelerating. For one thing, clinical trials typically involve more patients now than they did 10 years ago, probably because more drugs are targeted against chronic conditions. Only about one in four drugs that enter human studies is approved by the Food and Drug Administration. “Companies are reluctant to engage in risky front-end research if they’ll get only average returns,” says Henry G. Grabowski, a professor of economics at Duke University.

Grabowski notes that whereas “a few drugs make a lot of money, the average drug doesn’t give a good rate of return when you consider that you must cover ‘dry holes’ and common costs.” Only about three out of 10 drugs cover their development costs after taxes, he estimates. “About 55 percent of the profits come from 10 percent of the drugs,” Scherer adds. “If a committee were setting prices, it would target the top 10 percent. If you cut those profits in half, industry profitability would go to substantially below normal.”

Opponents of price controls point to

Clipper Runs Aground

Everyone seems to be listening in these days: tabloids regale readers with the cellular telephone intimacies of the British royal family, and more sober articles on the business pages tell how companies—or governments—devote resources to “signals intelligence” for commercial gain. So the Clinton administration might have thought it was doing everyone a favor in April when it proposed a new standard for encryption chips, developed with the aid of none other than the National Security Agency (NSA).

Instead the administration met with outrage. Along with the message, Clipper, as the chip is named, sends out a string of bits called a law enforcement field. Its purpose is to enable the police and the Federal Bureau of Investigation to decode conversations that they wiretap pursuant to court order. In addition, the chip’s encryption algorithm, known only as Skipjack, is classified. Thus, only a small cadre of cryptographic experts would be able to study it to determine whether or not it was indeed secure.

Early in June the administration abandoned its plan to rush Clipper into the marketplace and extended its internal review of the policy issues raised by the chip until the end of the summer. This decision presumably also delays consideration of outlawing other encoding methods.

A peculiar coalition of civil libertarians and large software companies has formed to plead the cause of unregulated cryptography. Self-styled “cypherpunks” argue that the government has no more right to insist on a back door in secure telephones than it does to restrict the language or vocabulary used in telephone conversations on the grounds that dialect might hinder interpretation of wiretaps. Companies such as Microsoft, Lotus Development Corporation, Banker’s Trust and Digital Equipment Corporation are worried about the administration’s proposal because they believe it will hurt the U.S. in international competition.

Ilene Rosenthal, general counsel at the Software Publishers Association, which numbers 1,000 companies among its members, points out that telephones containing the NSA chip would be subject to export controls because cryptographic equipment is considered “munitions” under U.S. law. This bureaucratic restraint could force U.S. manufacturers of secure telephones to develop entirely different product lines for the domestic and international markets. Indeed, she says, even if the State Department did license Clipper for widespread export, it is doubtful whether any foreign government or company would buy a system to which the U.S. literally held the keys.

Rosenthal contends that the U.S. has lost control of cryptography, citing more than 100 different brands of strong encoding software sold by non-U.S. companies. Indeed, discussion groups on the Internet computer network have been buzzing with plans for a do-it-yourself secure telephone that requires little more than a personal computer, a $200 modem and a microphone.

It is not clear whether the administration will abandon its attempt to stuff the unregulated-cryptography genie back in the bottle. There are already 10,000 Clipper-equipped telephones on order for government use, and Jan Kosko of the National Institute of Science and Technology says the plan for the standard “is being advanced as fast as we can move it.” Nine (thus far unidentified) cryptographers have been invited to review the algorithm, and Kosko reports that decoding equipment is in the works for “law enforcement and other government agencies that feel they need it.” The Justice Department is busy evaluating proposals for the “key escrow agents” that are supposed to prevent the police and the FBI from listening in on conversations without a warrant.

Some companies, however, are less concerned. They hope for enormous sales once privacy issues are resolved. AT&T, for example, announced its intention to sell Clipper-based telephone scramblers the same day that the chip was made public. “What the standard is,” says spokesman David Arneke, “is less important than having a standard that all manufacturers can build to.” —*Paul Wallich* the biotechnology industry, which is seeking to develop innovative types of therapeutics but depends on the confidence of investors. "The opportunity to raise capital by small biotechnology firms is now virtually nonexistent," George Rathmann, head of Icos Corporation, told Pryor's committee in February.

Even established pharmaceutical manufacturers have suffered drops in their stock market value because of fears about price controls. Recently, 17 members of the PMA have, in an effort to fend off mandatory controls, volunteered to restrict price increases on existing products to the rate of consumer inflation. In exchange, they seek the freedom to continue setting their own prices for new drugs. They promise to police their scheme and give it teeth if they obtain permission to discuss prices from the Justice Department. The industry says the permission, called a Business Review Letter, is needed to avoid possible antitrust charges.

The industry's voluntary effort seems unlikely to satisfy Pryor and his congressional supporters. Pryor points out that the voluntary limits apply to weighted averages of product ranges. Manufacturers could consequently reduce prices to bulk buyers while sharply increasing those for individuals.

Nor is Pryor convinced that research will suffer if prices are limited. He notes that 22.5 percent of drug manufacturers' revenues go to advertising and sales, substantially more than the amount that goes to research and development. And he makes hay out of the fact that U.S. companies often charge significantly more for their products in the U.S. than they do in other countries. In Canada, which has a Patented Medicine Prices Review Board, drugs cost 32 percent less than in the U.S. Companies, such as Merck, who say they strive for equal prices in different countries constitute an exception.

Pryor objects to U.S. consumers' being charged more because the drug companies benefit from several special tax breaks above the normal research write-off. Pryor has proposed limiting a tax break that gave tax savings of $1.5 billion last year to pharmaceutical companies with manufacturing operations in Puerto Rico. And the profitable links that companies can form with federal researchers through what are known as Cooperative Research and Development Agreements are also likely to be examined before the end of the year.

Drugs were always a high-risk, high-return business. Whether the U.S. industry will play well under different rules is a question that health care reformers have to consider. —*Tim Beardsley*

Blue Films

Cheap new coatings trap solar energy as color

One summer's day at Princeton University in 1991 a funny thing happened on the way to Guyot Hall. As Lori A. Vermeulen, a doctoral student in chemistry, carried samples of an organic/inorganic composite across campus, the white powdery material turned blue. "This was surprising," Vermeulen recalls. "When these compounds decompose, they turn yellow, brown, red—but they don't turn blue."

Back in the laboratory Vermeulen and assistant professor Mark E. Thompson got another surprise. They found that the material, zirconium phosphonate viologen chloride (ZPVC), changed color not because it was decomposing but because it was absorbing—and storing—ultraviolet light energy.

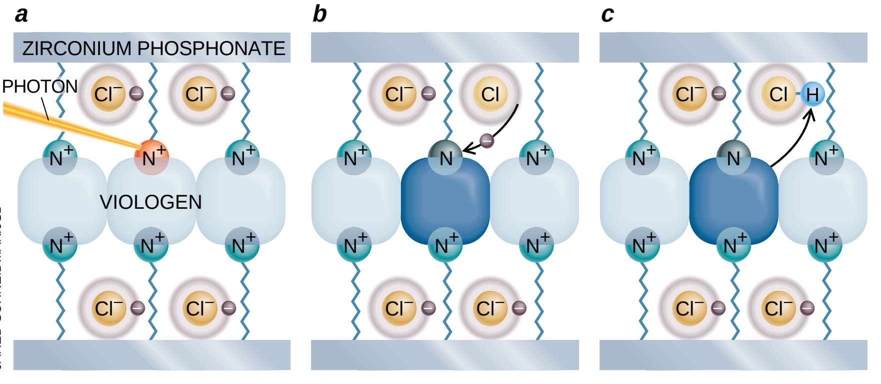
The discovery may have been right on the money. The two chemists had been studying ZPVC and related compounds with the idea of eventually building a chemical system to mimic the photosynthetic processes that convert light into chemical energy. They assumed that ZPVC would at best be just one piece of such a system, so they had not exposed it to ultraviolet light in the lab. Experiments since have shown that the blue color—and the stored energy it represents—persists for months. If working devices can be made from the material, it may become a low-cost alternative to silicon solar cells. But instead of producing an electric current, these cells would work as solar-powered catalysts.

Many compounds can convert light to an electrical charge, but very few can hold that charge. When an atom absorbs a photon, the added energy excites one of its electrons, often to such a frenzy that the electron leaps onto another atom. This electron transfer creates ions, or radicals, out of the donor and acceptor atoms. Radicals can do all kinds of useful things: split water into hydrogen and oxygen, for instance.

If charge transfer were this simple, however, artificial photosynthesis would be commonplace. Unfortunately, the transaction is nearly always rejected: the electron bounces back to its donor, and the energy dissipates. Researchers have spent years devising clever systems—using zeolites, porous glasses, fullerenes and other exotic materials—to hold the donor and acceptor atoms at just the right distance to allow electron transfer but prevent the bounce. When Vermeulen and Thompson first published their findings in *Nature* last year, they hypothesized that the rigid lattice created by the zirconium atoms in ZPVC did this naturally.

"Those proposals are probably wrong," Thompson now admits. ZPVC does undoubtedly hold on to its charge fiercely: two years later Vermeulen's initial sample is still decidedly blue. But further experiments have convinced Thompson that "there are at least two different processes going on. One is reversible, but another is irreversible," meaning that every time a bit of ZPVC converts light to charge (and changes from white to blue), it gets used up [*see diagram below*]. The

ROBERT PROCHNOW



JARED SCHNEIDMAN/JSD

ZPVC CAPTURES LIGHT and stores it as chemical energy. A photon hitting white ZPVC* (a, top left) *excites a viologen group, which then steals an electron from a nearby chloride ion* (b) *and turns blue* (top right). *The chlorine seizes a hydrogen atom* (c). *The color and energy persist for months, but the process is irreversible.

reversible process "is still a big question mark," Vermeulen says, "but an exciting one, because a reversible mechanism is exactly what you want" to make solar catalysts.

Equipped with a full understanding of at least part of ZPVC's inner workings, the chemists are now beginning to tinker with them. Their first challenge is to make a porous version through which they can pump chemicals that will react with the radicals to extract the stored energy. By replacing the viologen with smaller chemical groups, the researchers have opened up spaces in the matrix.

They have also formed films of ZPVC, 20 layers thick, on glass and other substrates. "Because there are so few layers, oxygen can get in close enough to remove the electron. So as soon as we expose the films to air, they go back to their original color," Vermeulen notes. "That's actually a good thing," she adds, "because you want to be able to get something in there to react."

The films also seem to respond better than powdered ZPVC to light, turning darker blue and holding more of their color over time (so long as oxygen is kept at bay). Thompson has found that 15 percent of the photons absorbed by the films lead to charge separation, a respectable number by the standards of artificial photosynthesis.

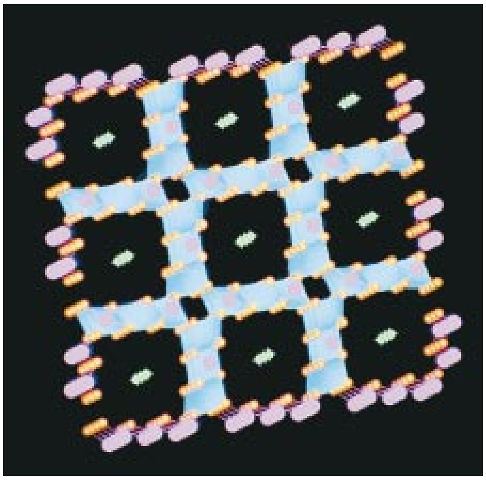
By the standards of solar cells, however, ZPVC has a long way to go. Its finicky electrons ignore all but ultraviolet light, and, as Vermeulen points out, "there isn't much light outside at that wavelength." Adding ruthenium or other elements might sensitize ZPVC to visible light, "but that opens a whole new can of worms," Thompson says. "We plan to stay in UV for a while."

The scientists are circumspect about commercial possibilities. The material itself is "dirt cheap" and "really easy" to synthesize, Vermeulen says. And if the mysterious reversible pathway can be understood and harnessed, "there's still a chance we may be able to use these compounds as solar harvesting agents," Thompson reflects.

He and his graduate student intend to find out, anyway. Thompson has applied to the Department of Energy for a grant, and, Vermeulen says, "we also have some other industries and individuals interested in funding us." Princeton, meanwhile, is rushing to put together a patent application so that negotiations can begin. "This technology is not at the point of commercialization yet," Thompson emphasizes, "but there are people who want to be involved now so that they will be involved later." Who knows? Someday it may actually be respectable to produce blue films. *—W. Wayt Gibbs*

A Gem of a Catalyst

Todorokite (pronounced "toh-doh-ROH-kite") rolls off the tongue like the name of a gemstone. It is really just a form of common manganese oxide. But to some chemists, it has all the panache of a precious mineral. They have for the first time been able to synthesize stable forms of the substance, which behaves in a manner similar to zeolites—industrially important compounds used in filtration and catalysis.



STEVEN L. SUIB

***TUNNELS** of synthetic todorokite, made of manganese* **(pink)** *and oxygen* **(orange)** *atoms, have each adsorbed a calcium ion* **(green)** *in this computer-generated image.*

In nature, todorokites belong to a family of related compounds that are found on the ocean floor and in soils, appearing to the naked eye as small, black pellets. Like zeolites, natural substances of aluminum, silicon and oxygen that are now widely synthesized, todorokites contain a network of tunnels that can adsorb other compounds and later release them. "They act as molecular spaghetti strainers," says Steven L. Suib, a chemist at the University of Connecticut, who headed the team that made the substance.

In the 1960s investigators suspected that todorokite could be commercially practical because the mineral can exchange ions and contains large tunnel diameters, about 6.9 angstroms square (0.69 nanometer). "They found that the natural materials were extremely good oxidation catalysts," Suib says. The problem was that the natural materials were difficult to purify and study. Attempts to make todorokite in the laboratory met with limited success. The early compounds were not thermally stable at temperatures required in commercial reactions.

Now a collaboration between researchers at the University of Connecticut and Chi-Lin O'Young of the Texaco Research Center in Beacon, N.Y., has taken a major step in developing commercially viable synthetic todorokite. Using an idea first proposed by Joe B. Dixon, a mineralogist at Texas A&M University, and his colleagues, the team was able to produce the mineral in a pure, stable form. The researchers concluded they had todorokite, in part by allowing their compound to adsorb hydrocarbon molecules of different sizes. The largest particle adsorbed had a diameter of about 6.9 angstroms—in agreement with the natural form. The synthetic molecular sieves were stable to about 500 degrees Celsius.

Although most zeolites are stable at higher temperatures, Suib thinks the synthetic version has important potential applications. Many types of ions and metals are more readily adsorbed by todorokite than by zeolite. Furthermore, todorokite, unlike zeolites, is electrically conductive, so it could be exploited as electrochemical sensors or batteries. Indeed, Argonne National Laboratory is looking into the mineral as a possible basis for a new type of rechargeable lithium battery. The todorokite tunnels would act as a housing for the charge. "Lithium would go in during the charge and back out during the discharge," Suib says. A few battery manufacturers have also inquired about the synthetic material.

Todorokite could also be useful in environmental engineering to filter out heavy metals. Ions of zinc, cobalt, lead and nickel seem to be especially fond of manganese oxide tunnels. "Todorokite and other manganese oxides all tend to be rather reactive and to occur deep in the soil," Dixon says. "These highly selective oxides must play some kind of role when the ions dissolved in rainwater leach into the soil." The synthetic versions, being pure, should enable mineralogists to characterize more fully the chemical role of the natural products. What more could Tiffany ask? *—Philip Yam*

Will Deregulation Save the Banks?

During the past few years, the commercial banking industry has been riding on the brink of a debacle even larger than the savings and loan disaster of the mid-1980s. Although economic conditions appear to be giving banks some breathing room, proposals for strengthening the underpinnings of the system have evoked disagreements among economists. Indeed, some claim that federal deposit insurance, long considered essential to preventing bank panics, may cost taxpayers more than it is worth, unless the U.S. makes fundamental changes in the structure of its banking institutions.

At the end of 1992 there were 950 "troubled" banks, with $500 billion in assets, according to John P. LaWare of the Federal Reserve Bank Board of Governors. And the reserves of the Federal Deposit Insurance Corporation (FDIC), which stands behind deposits up to $100,000, were nearly depleted from closing or reorganizing shaky commercial banks over the past few years.

Since January, the rate of failure has slowed, thanks to low interest rates and the slight improvement in the economy at the end of 1992. That, along with a $70-billion loan from the U.S. Treasury to the FDIC, has averted an immediate crisis. Nevertheless, about 800 banks remain on LaWare's list. An economic downturn could squeeze banks further by increasing default rates on loans and deflating the value of the assets backing them. And a surge in interest rates would narrow the spread between what banks must pay depositors for funds and the return they earn on their loans.

The precarious state of the industry has once again been accompanied by calls for deregulation. According to James R. Barth, an economist at Auburn University, the combination of government-insured deposits and laws that restrict bank diversification has produced serious flaws in the system. Barth and his colleagues R. Dan Brumbaugh and Robert E. Litan were commissioned by Congress to do a study of the industry, recently published in book form as *The Future of American Banking.*

One of their main targets is federal deposit insurance, which was instituted in the 1930s to end the "runs on the bank," or panics, that took place in the U.S. nearly every 20 years from 1857 to 1929. These economic meltdowns occurred when the failure of one bank precipitated instability in others and stirred anxiety among depositors. The subsequent rush to withdraw savings caused additional failures. The money supply dropped sharply, restricting commerce, and credit for long-term loans dried up as survivors attempted to avoid risk.

Although insurance put a stop to panics, Barth and others argue that the FDIC has an insidious downside: it encourages banks to undertake more dangerous investments. It is a classic example of what economists call "moral hazard"—insurance inducing people to behave in a manner that increases the likelihood of the event that is insured against. During the high-flying 1980s, commercial banks invested so heavily in loans for real estate, less developed countries and corporate restructuring, they contend, because their accounts were insured against loss. Depositors had no incentive to discipline poorly managed banks by withdrawing funds.

Some experts claim that federal deposit insurance makes failures only worse.

Furthermore, regulators had an incentive to underestimate bank ills. Reports of large numbers of insolvencies suggest that regulators are not doing their job, and it costs the FDIC money in the short run (even though it might save in the long run).

The result is "adverse selection": poorly managed banks and their depositors avoid the consequences of imprudent investments. Meanwhile institutions with safer assets earn lower returns but must pay the same interest rate as do their riskier colleagues, to compete for funds. As an added punishment, well-managed banks end up having to pay higher insurance premiums to cover banks that go under.

Barth—along with the Board of Governors of the Federal Reserve Bank—argues that the legal restrictions barring banks from operating nationwide and from offering a full range of financial services also exacerbate the industry's problems. Imposed by the Glass-Steagall Act in 1933, these rules were intended to stabilize the banking system by barring banks from underwriting stocks, holding equity in corporations or doing business in more than one state.

These prohibitions against universal banking have limited the ability of commercial banks to adapt to market changes and to increased international competition. Nor have they been able to compete with nonbank financial institutions, such as mutual funds, which now offer services almost indistinguishable from the conventional checking or savings account.

Some economists speculate that it might be quite beneficial to allow banks to underwrite and market securities. If banks owned equity in the firms they loaned money to, they might even influence management in beneficial ways (as they reportedly do in other countries, such as Japan). Although the government has loosened Glass-Steagall restrictions in the past few years, Barth and some other economists call for the elimination of all of them. Universal banks tend to be the rule rather than the exception elsewhere.

Such deregulation, however, would make it difficult to preserve deposit insurance, which may be guarding against economic cataclysm. Deposit insurance is fine when banks make conservative investments and failures are rare, but it is not rational for taxpayers to underwrite the risks unless they can also restrict the kinds of investments that banks take on.

Frederick S. Mishkin of Columbia University warns that "widening the scope of permissible financial activities will make it easier for some banks to engage in high-risk strategies." Moreover, he notes, a universal financial services company "would have strong incentives to tap its bank as a source of funds" if it ran into trouble in other areas. Because of this peril, Mishkin doubts that universal banking will be a reality in the U.S. anytime soon. "It will be too dangerous, unless things change pretty dramatically," he says. —*Judith Fields and Paul Wallich*

JUDITH FIELDS is a fellow at the Jerome Levy Economics Institute at Bard College in Annandale-on-Hudson, N.Y.

Circuits That Get Chaos in Sync

Chaos is not always so chaotic. In some sense, it can be predictable: two systems can be designed so that one exhibits exactly the same chaotic behavior as the other. In other words, the systems would be synchronized. Such devices might be useful for encrypted communications. For example, one of the systems could conceal a message within the chaotic signal. Only someone who possesses the second system would be able to decode the transmission, by subtracting the chaotic signal and leaving behind the message [see "Mastering Chaos," by William L. Ditto and Louis M. Pecora, page 78].

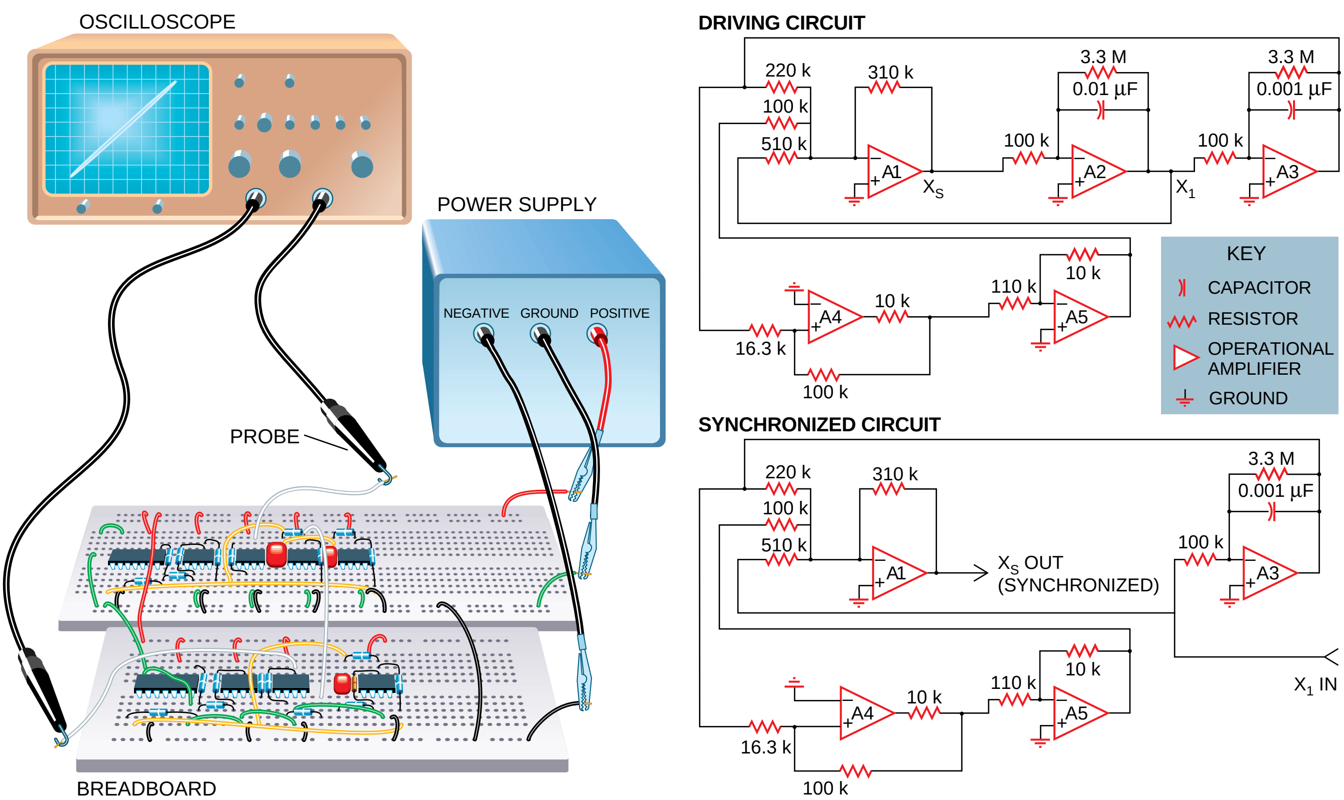
Louis M. Pecora of the U.S. Naval Research Laboratory first came up with the idea of synchronized chaotic systems. He and one of us (Carroll) used a computer simulation to show that such a phenomenon can exist. The next step was to demonstrate the idea using physical systems—specifically, electrical circuits, which are accessible and inexpensive. The first circuit that Carroll built to display synchronized chaos was based on a design devised by Robert Newcomb of the University of Maryland.

Although the circuits described here are simplified versions of Newcomb's, experience with circuit assembly might be helpful. A good introduction to chaos in electrical circuits, using only a diode, an inductor and a resistor, appeared in this column last year [see "How to Generate Chaos at Home," conducted by Douglas Smith; SCIENTIFIC AMERICAN, January 1992].

Basically, the setup will consist of two circuits: a driving circuit and a synchronized circuit [*see illustration below*]. The two are identical except for an important component missing in the synchronized circuit. The two circuits are connected at a single point. The chaotic output of the synchronized circuit will match that generated by the driving circuit if both circuits are correctly built.

Before buying the electronics parts and constructing the two circuits, make sure you have an oscilloscope, breadboards and a power supply that can deliver 12 to 15 volts of direct current. (Breadboards are thin sheets of plastic that have holes to accommodate electronic components; purchase ones that will hold the positive, negative and ground terminals from the power supply.) The power supply and breadboards

JOSEPH NEFF and THOMAS L. CARROLL devised the experiments described here by adapting Carroll's design of the first electronic circuits to exhibit synchronized chaos. Neff, who built the circuits during his senior year at the College of Wooster, recently graduated and is currently pursuing graduate studies in physics at the Georgia Institute of Technology. He programs computers to investigate chaos. Carroll, who received his Ph.D. in 1987 from the University of Illinois, works at the U.S. Naval Research Laboratory. He looks for applications of chaos and has patents pending on some chaotically synchronized circuit designs.



CHAOTIC CIRCUITS produce the same output. They draw current from a 12-volt power supply and are probed by an oscilloscope (left). *The values for the components used should match those shown in the schematics* (right) *to within about 1 percent.*

are cheap and easy to come by. Oscilloscopes, unfortunately, are expensive; they start at about $500. You can sometimes buy a cheap, used oscilloscope at a "hamfest"—a flea market for ham radio enthusiasts. If you do not wish to purchase an oscilloscope, you might be able to get access to one at a local college laboratory. The device should provide two channels and be able to plot the input to the channels against each other.

For the circuits themselves, you will need various resistors, capacitors and integrated chips called operational amplifiers (op-amps). A list of such items and the minimum quantity you need appears below. We recommend buying more than the minimum, because it is easy to make connection errors that can burn out the components, especially the op-amps.

Because the two circuits must be as similar as possible, it is essential to use resistors and capacitors that have high tolerances. Look for resistors rated to be accurate within 1 percent and capacitors made from polypropylene, which do not leak much current. All but two of the op-amps are generic type 741. The exceptions, labeled "A4" on the schematic [*see right illustration on opposite page*], are so-called high-frequency uncompensated op-amps. The particular ones used here were type NE5539N. These two are the most critical components of the circuits, as they are ultimately responsible for keeping the circuits synchronized.

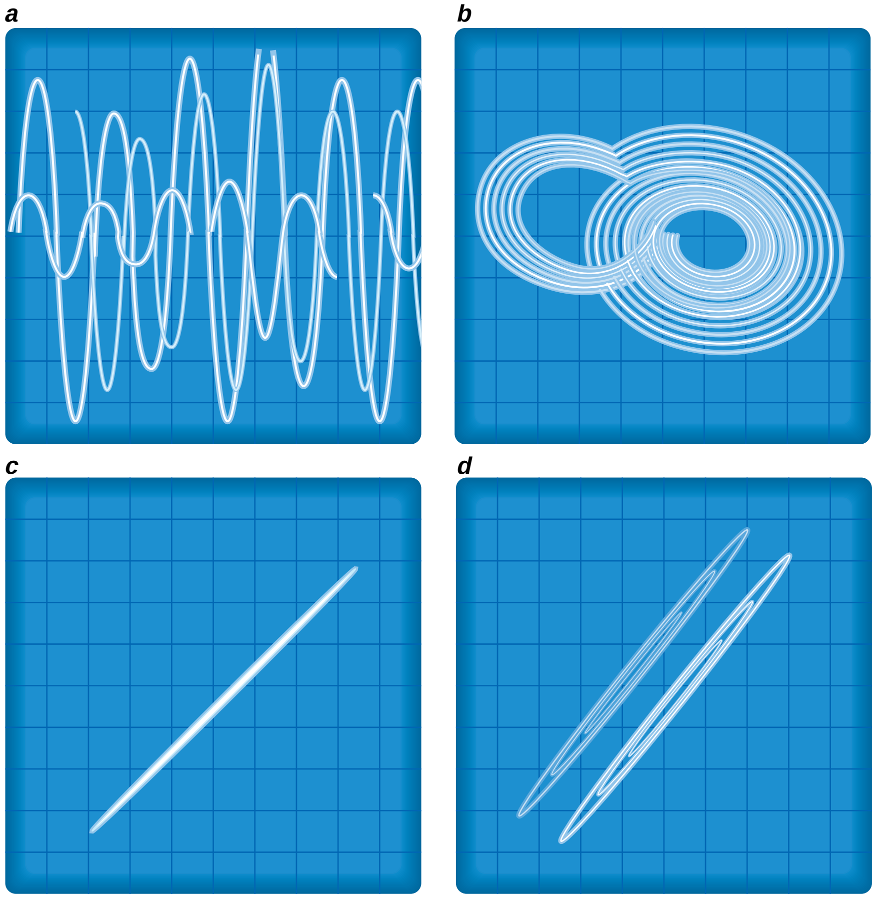
The components we used were purchased from Digi-Key Corporation in Thief River Falls, Minn. The total price was less than $30. When ordering, remember to ask for the pin diagrams for the op-amps, which contain several connections that all look alike. The specification sheet describes the configuration of the pins so that you will know which ones to use.

To decrease the possibility of wiring error, it is best to lay out the circuits according to the schematics before connecting them. Start with the op-amps, using one breadboard for each circuit. Observe that the synchronizing circuit is identical to the driving circuit, minus an op-amp and its ancillary components.

The next step is to make the power connections (be sure the power supply is unplugged before beginning). Only the op-amps draw current. Attach wires from the positive, negative and ground terminals to separate rows that run along the top and bottom of the breadboard. These rows are dedicated solely to providing voltage to the electronic components. Use the op-amp spec sheet to determine the appropriate connections from the pins to the rows. The ground wires should be as short as possible because long ground connections can pick up noise, which would keep the circuits from synchronizing.

After hooking up the power supply, begin connecting the components systematically. Although the schematics are simple, it is easy to create a mess. Short, color-coded wires make the circuit convenient to read and check over. Take several different visual walks through the circuits to verify that all of the connections are correct. Writing down the color codes of the resistors next to their values on the schematic may speed up this process.

Once you are satisfied that your circuits are correct, connect the synchronizing circuit to the driving circuit at the points marked "X_1" on the schematic. Notice that as a result of the connection, the two circuits have the same nonlinear driving component (the subsystem that uses the A2 op-amp). Connect each channel of the oscilloscope to the points marked "X_S" on the schematic. Each such point is located just after the A1 op-amp. Make sure that the circuits and the oscilloscope all have the same ground—that is, the grounding wire attached to the oscilloscope's probes should be hooked up to the same ground line as the circuit.



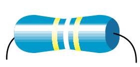
***OSCILLOSCOPE PATTERNS typically found in working synchronized circuits are shown above. The chaotic output from one circuit is revealed by a plot of output voltage over time* (a)*; several sweeps are shown. Plotting the output from A2 against that from A3 yields a chaotic attractor* (b)*. Comparing the output of both circuits at X_s shows the circuits are perfectly synchronized* (c) *or partially so* (d).**

COMPONENT SHOPPING LIST

CAPACITORS

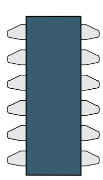
- 2 - 0.001 µF
- 1 - 0.01 µF

RESISTORS



- 4 - 10 k
- 2 - 16.3 k
- 7 - 100 k
- 2 - 110 k
- 2 - 220 k
- 2 - 310 k
- 2 - 510 k
- 3 - 3.3 M

OP-AMPS



- 7 - TYPE 741
- 2 - TYPE NE5539N

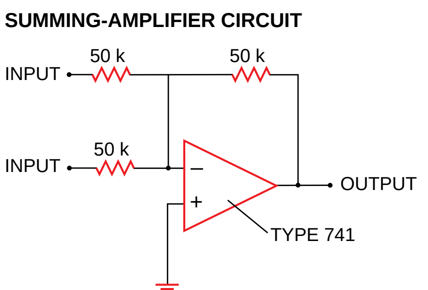
µF = MICROFARAD
k = KILO-OHM
M = MEGA-OHM

Check the circuits several times before plugging in the power. If you burn out one of the op-amps, you can plan to spend at least a few minutes trying to figure out which component is broken and replacing it.

You can verify that the circuits are producing chaos by displaying information from one channel only. If that circuit's output is chaotic, the oscilloscope image will not remain stable on any setting. Instead it produces a pattern similar to a sine wave that rapidly changes in amplitude.

A detailed explanation of why chaos arrives in the driving circuit is rather complex. Briefly, the two integrating op-amp components (the op-amps that have capacitors, labeled "A2" and "A3") are connected in a loop to generate a sine wave. The op-amp components A4 and A5 amplify this sine wave, causing its amplitude to increase exponentially over time. Once the signal reaches a certain amplitude, the high-frequency op-amp (A1) brings the amplitude to zero and starts the process all over again. Chaos creeps in during the switching, when A1 resets the circuit.

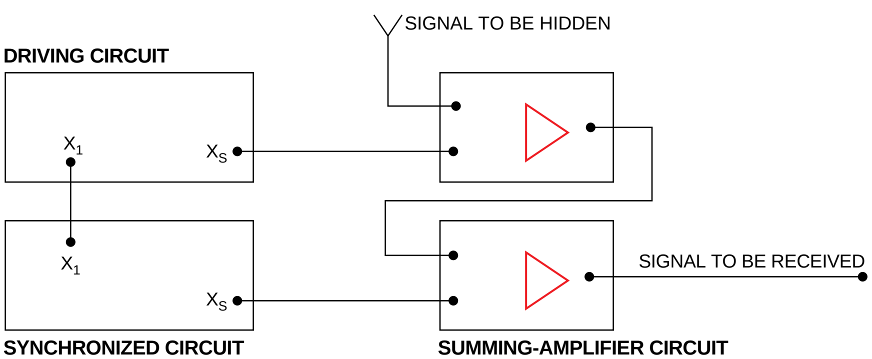
You can see if your circuits are producing synchronized chaos by plotting the output of one circuit at X_S with that of the other. If the circuits are perfectly synchronized, the oscilloscope will produce a straight line angled at 45 degrees. Because of the switching, the synchronization will probably not be exact. Two broadened, parallel lines may be the result instead.



Do not be disappointed if the circuits fail to synchronize on the first try. Check for faulty connections. The circuits are rather fussy, so any poorly made connections could wreck the synchronicity. A faulty op-amp could be the culprit. You might find that both circuits generate chaos but are not synchronized. There are two possible explanations. One, components that are supposed to be nearly the same may differ too much; for example, two resistors designated as 1,000 ohms may turn out to be 950 and 1,050 ohms. Two, the response time of the high-frequency op-amps may be insufficient, in which case the op-amps should be replaced.

Once the synchronizing chaotic circuits are working properly, you can create the form of message encryption envisioned by Pecora and Carroll. You will need two more op-amps and several large (high-impedance) resistors. With these components, you can easily build two "summing-amplifier" circuits. Such circuits add two signals together and then invert them. These two circuits are connected to the driving and synchronizing circuits [*see illustration below*]. The message we encoded was a simple sine wave created by a function generator. The sine wave is fed to one of the summing-amplifier circuits, which combines the sine wave with the chaos generated by the driving circuit and inverts the total signal. Amid the chaos, the sine wave is almost impossible to decipher.

The synchronizing circuit, however, can readily extract the message from the chaos. Because the synchronizing circuit produces the same chaotic pattern as the driving circuit, all one has to do is add the synchronizing circuit's chaotic signal to the driving circuit's total (inverted) signal. What is left is the message. The second summing-amplifier circuit does this job and reinverts the sine wave to its original form.



***SUMMING-AMPLIFIER CIRCUIT* (top) *marries an input signal with a chaotic signal and then inverts the combination. Two such circuits, when connected to the synchronized chaotic circuits, produce a form of encrypted communication* (bottom).**

The summing-amplifier circuits may cause the entire system to behave quite erratically; the image on the oscilloscope may blow up. The problem is that the external devices drain some current and thereby jolt the circuits from one so-called chaotic attractor—a pattern around which the chaotic signals tend to settle—to another.

Large resistors help to alleviate the problem by preventing current drain. We used resistors of 60 kilo-ohms and one mega-ohm, although 50 kilo-ohms will probably do. Another way to get the circuit back to a single attractor is to short out an op-amp periodically. Briefly touch a grounded wire to the output of one of the op-amps (preferably the ones with capacitors). This process randomly resets the state of the circuit. The op-amps will not be damaged so long as the shorting is brief. Turning the power supply off and then on again will also work.

To hear the message, you will need to amplify the output. Op-amps do not deliver much current (and the high-impedance resistors do not help matters). A stereo system is a nice solution. Alternatively, a basic electronics handbook will detail how to build simple amplifying circuits to drive a small speaker.

You might try to encode signals more complex than sine waves. You can send audible messages by using a microphone, which converts sound waves into concealable electrical signals. Or you can try sending Morse code: each "click" sends a constant voltage signal. On the other end, you can wire up a small light bulb or a light-emitting diode (known as an LED), which should flash in unison with the clicks if the circuits are working properly.

You can also demonstrate the effects of attractors in chaos by plotting the output of one of the integrating op-amps versus its neighboring op-amp. Such an input produces a spectacular pulsating spiral pattern and is evidence for a single attractor. Feel free to change some of the components to see what happens. Remember, you are trying to create chaos.

FURTHER READING

NONLINEAR CIRCUITS HANDBOOK. Analog Devices Engineering Staff. Analog Devices, Norwood, Mass., 1976.

DYNAMICS: THE GEOMETRY OF BEHAVIOR, Parts 1–4. Ralph H. Abraham and Christopher D. Shaw. Aerial Press, 1985.

INTRODUCTORY ELECTRONICS. Robert E. Simpson. Allyn and Bacon, 1987.

THE ART OF ELECTRONICS. Paul Horowitz and Winfield Hill. Cambridge University Press, 1991.

BOOK REVIEWS by Philip Morrison

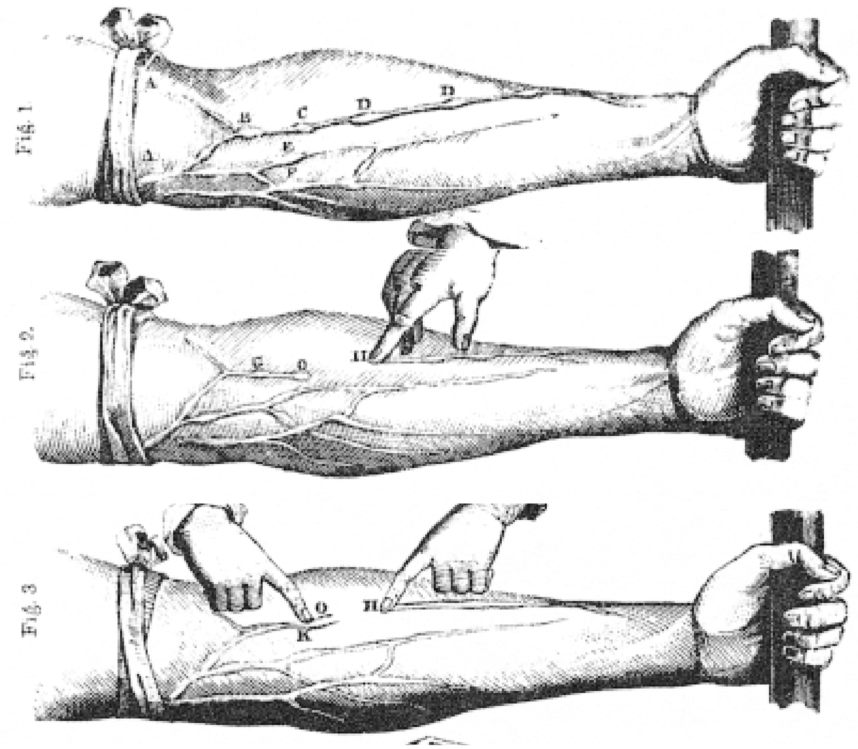
Matters of the Heart

VITAL CIRCUITS: ON PUMPS, PIPES, AND THE WORKINGS OF CIRCULATORY SYSTEMS, by Steven Vogel. Oxford University Press, 1992 ($24.95; paperbound, $12.95).

An adept at fluid biomechanics—from airflow through prairie dog burrows to nutrient seawater past arrays of hungry clams—wide-ranging Duke physiologist Steven Vogel found his mind wonderfully concentrated not long ago. He was loading hand-sawed firewood into his pickup one autumn day; he strained to lift and hurl that last unsplit chunk to the top of the load. The effort cost him 10 days in the hospital, another repentant from the Valsalva maneuver, a few stressful seconds of exhaling against a tight-closed throat.

In this maneuver the air pressure inside the thorax rises substantially as the body stiffens like an inflated tire. The action does take load off the backbone, so that it comes naturally enough. But since 1707 it has been recognized as risky, even dangerous. The increased internal pressure must be matched by that of the venous blood, so that for some seconds the flow of blood out and back to the lungs dwindles. The working heart, "that consummately aerobic muscle," is starved of oxygen just at the peak of its demand for power. "If you're prone to heart trouble, the stunt is likely to bring it on." Most urban snowfalls offer plausible epidemiological evidence, a flurry of heart attacks among middle-aged shovelers, their breath too often held during quick, unusual exertions. Cheerfully devoted to his own rehabilitation, Vogel is no longer the pudgy meat eater he was. No longer does he hand-saw his own wood; the fun has gone out of it.

We readers, too, have benefited from that last incautious heave. In an earlier book, this deft writer on all of flow biology excused himself from addressing human circulation. He saw that intricate literature as prolix and narrowly medical. Now he faces up to the task at the simplest, most needed level, with little Greek and not even one equal sign. (One misses some helpful graphs.) That this topic is at once both intimate and fateful does rivet us, and Vogel's originality and wit give his engaging book



ACTION OF VENOUS VALVES was demonstrated early in the 17th century by William Harvey. This illustration is an 1847 redrawing of Harvey's original.

both the freshness and the steady beat of suspense the more methodical, jargon-beset textbooks lack. He has not put aside his comparative methods, treating spiders, mudflat worms, thirsty giraffes, even trees, to enrich the central narrative of those three billion heartbeats that count out our lives. Tough-minded attention to business is leavened by pointed and candid reflections on the methods and purposes of science well performed.

Scale is determining. In the small, the essential transfers of oxygen, heat and subtler molecular messages of control are mere random diffusion. Directed, convective flow is obligatory mainly for us larger, warmer creatures. The biology of many-celled animals proceeds on two disparate scales: "There are about as many molecules in a cell as there are cells…in a cat." Yet even bulk flow can be powered at the cellular level by what Vogel terms "micropumps," tiny, beating cilial hairs that line some tube walls. The currents that nourish the filter feeders of the sea, the tidying flow of mucus out of the human trachea and, most abundantly, the flow of sap in trees are all under micropump drive, at some extra cost in power. The spotlight dwells on macropumped flow and the pressure differences that drive it through a remarkable adaptive network of compliant tubing, thick-walled arteries and slighter veins. Each is elegantly matched to its task; the capillary net links all that is between, some vessels so fine-drawn that the red cells must jostle past in single file.

The pumping human heart is as double as its symbol. (In this domain, language is rich with unintended metaphor.) Your two half-hearts, left and right, are two distinct pumps that beat as one, held in a single contractile housing spun of muscle and collagen. Readers are urged and guided to buy a real mammalian heart for direct examination, either untrimmed from the butcher or from lab suppliers at a variety of distancing ploys, a merely embalmed heart to one fully visible but untouchably cast in plastic. For our author, a heart is nothing to be wasted, but it cannot be naively cooked just like the skeletal muscle and fat of a leg of lamb; heart walls are low in fat, muscle fibers laced with rubbery fibrin and strong collagen, and only long, slow boiling in acidic solution makes them good eating. A highly spiced Madras recipe, founded on lemon and vinegar, is offered; you might look up a sauerbraten.

The two flow loops that the half-hearts power are in series; the right half-heart feeds spent blood fluid out to the lungs, to return to the larger pump at the left side. That one sends the aerated heart's blood, more crimson than rust, through a myriad of paths to bathe about every body cell. The outflow rejoins eventually to the right side of the heart. The series pumps are self-compensating; their beat is single, and the stroke volume adjusts itself without any extra mechanism because of the relation between stretch and tension intrinsic to muscle fiber. Extra input volume causes increased pressure and hence volume of output over some range of stretch. Alas, a weakened heart wall, outside of that normal range, may blow out instead: one form of heart attack.

Dr. William Harvey's classical arguments are here; after four centuries, they still persuade. Even his early rough quantification of output flow was as accurate as it needed to be to answer his seminal question. The aim of quantity in science is not mere maximum precision, but approximations reliable enough to argue from. The person who measures your blood pressure is careful to place the pressure cuff at the same height above the floor as is your heart. You can even practice a little deception: to alter your pressure reading, slouch down in the chair. Lowering the heart a few inches below the arm, when the arm is supported on a table, will reduce the readings by five or 10 points. Yet this old and noninvasive test is remarkably accurate for all its simplicity of means.

The stethoscope we all know is an even older wonder. The physician listens especially for troubling turbulence in flow, although only the largest blood vessel, the aorta, works at all close to the turbulent regime. Sometimes he hears a quiet murmur, sign of a trickle of backflow past some flawed seal in one of the four automatic check valves the vital pumps require. In these matters of the heart, quantity is everywhere.

An Uncertain Record

WEATHER CYCLES: REAL OR IMAGINARY? by William James Burroughs. Cambridge University Press, 1992 ($39.95).

Everywhere under heaven the weather is real enough; men and women by the hundred million fear hunger or feast merrily as each Indian monsoon fails or succeeds. It was Theophrastus, Aristotle's postdoc, who remarked that both ends of the lunar month were apt to be stormy, a weather cycle still applied in the forecasts of the *Farmer's Almanac.*

People have always tried to tease out some repetitive order hidden in the unending variations of weather. Since the Enlightenment they have pored over human records of weather the world over, as well as major "proxy data," like tree rings, sunspots and the sequenced samples of deep-sea ooze and polar ice. Much of all these is warily set out here in a well-written and judicious account of largely unresolved controversy, necessarily displayed on long, wiggly graphs and in numerical tables. Meant for general readers with real interest in the weather, the book draws perforce on some mathematics, quite clearly explained.

Long ago and far away, in the Precambrian rocks of Australia, thin layers of coarser sandstone and finer siltstone were laid down one on another. In a well-marked rhythm of changing thicknesses, they can be seen today on 200 feet of exposed rock face. The geology seems familiar; such layers in recent times record the fluctuating entry of sediments into a glacial lake when meltwater fills it every spring. Fine silt settled when the flow was slow, coarse sand when swift freshets could carry bigger grains. In the Australian rocks, 19,000 thin laminae are patently grouped between thicker ones, eight to 16 layers to a group, around an average of 12. (Photographs appeared in this magazine in August 1986.) Plots of the variations in thickness bear "an uncanny resemblance" to sunspot number counts since the time of Galileo, but here held in rock 700 million years old.

Was this the long-sought smoking gun that proves weather change follows the cycle of solar activity—12 years then, as it is 11 years now, between peaks? Later analysis raised real doubt: no smoke, maybe not even a gun. The sediments may not be annual layers but only daily tides in some salt lagoon, of no interest for climate. "Without this example the evidence of periodicities in the annual layers laid down in lake beds is pretty slim." Tree rings, too, give confusing signs; the few regularities are local at best and seem to shift over time.

There are two main procedures to sift such records. The simpler is just the peaks and troughs as measured. Comparing wavy patterns of temperatures or even of wine harvest dates over many years can convince the eye. Of course, there are measures of statistical significance to be calculated, especially after manipulation of the raw data to redress the distortions of all limited sampling. The remedy for many of these troubles is an old one: the famous device of harmonic analysis, able to match any shape whatever as a sum of simple wave shapes, by allowing all wavelengths. Your plot is a spectrum: how much of each frequency component is present. Strong peaks that tower over random fluctuations point to a periodic cause with the frequency of the component peak. The spectrum most used today is called the power spectrum, easily computable, which uses the squared amount of each component in analogy to classical wave physics. The aim is prediction by an inverse computation that combines only the waves that build major peaks. It goes with the flow of our number-crushing age.

But a hurricane warning flag is up. Real complex systems possess feedback and respond disproportionately to trigger effects of certain small changes. They admit some chaotic behavior, typical of real weather. Spectral peaks can appear at twice the driving frequency, or at half of it, or at sums and differences of major causal frequencies. Important peaks may even come and go as time passes.

One experiment is gloomily clear. A group at the University of Texas at Austin set up turbulent flow within a fixed cylinder and an inner rotating one. They measured the speeds in the flow by measuring laser reflection from suspended metal particles and reported the wiggly curves they saw for faster and faster rotations. Next to the data they show the power spectrum. For low speeds, one spectral peak dominates, related to the rotation. A little faster, and you see an entire skyline, although the periodic input, rotation, has simply smoothly increased: typical chaos.

In 1990 two meteorologists at the University of Reading prepared a simple but clearly nonlinear numerical model for world weather. It had no external periods they could find, save the yearly succession of seasons. A century of the model's run made a fine power spectra. The one-year peak was sharp and high, but there were plenty of longer ones, a 12-year one above them all, although there was no sunspot cycle modeled or any other external period. Power spectrum agreement does not of itself demonstrate causality or even allow satisfying prediction.

Just what cyclic behavior is causal? Physical understanding is necessary, but even that might not be sufficient. We see good signs of quite a few periods in the weather: biannual shifts of stratospheric wind and surface temperatures, the recurrent linkage of tropical and subtropical sea temperature and weather along a lengthy path all the way from India to Chile, and much more. Ice Age weather sampled by deep cores does match some power peaks in the slow, periodic wandering of earth orbit and axial tilt in response to the gravitational pull of the moon and the planets. Burroughs, an informed critic, does accept that broad result, although it does not appear to be beyond doubt.

The biggest prize is unclaimed: How much do changes in global temperature reflect natural causes? Surely the enthusiasts for order have often been too confident after some limited success; it goes with all predictive territory. Sum up by saying that weather cycles contain both real and imaginary parts. The mathematicians, having quite another meaning in mind, say that such a quantity is complex. The wordplay seems apt for a book whose clarity and breadth of vision set it apart.

What We Seek in the Sky

CONVERSING WITH THE PLANETS: HOW SCIENCE AND MYTH INVENTED THE COSMOS, by Anthony Aveni. Times Books, 1992 ($21).

Some 20 years back Anthony Aveni and his Colgate astronomy students were detained for prowling by night around the great ruined city just north of the Valley of Mexico. They intended nothing more sinister than fixing the orientation of the high pyramid there by surveying the stars. Their escapade was worthwhile: that structure faces the direction of sunrise on two wonderful days—only in the tropics—when at noonday the sun passes straight overhead. His insight into celestial order as marked by the astronomers of Teotihuacán, so old that it was lost in time even when the Aztecs came first to the Valley, soon drew Aveni away from the telescopes of Kitt Peak for good. He has since become one of the most productive scholars of pre-Columbian astronomy. This informal volume is his readable, popular, empathic synthesis of the ancient search for celestial order in human context.

Although Aveni admires both unity and diversity within human discoveries, a reader notices a sharp list toward differences; by now the author is more anthropologist than astronomer. His examples span both the Old World and the New, mostly seen through one "primary lens," the bright wanderer we call Venus. (Planets rule here; the star Sirius does not even make the index.) In seven readable chapters, illustrated with plenty of sky deities and their glyphs, the author presents an appraisal of what men and women have sought in the sky. For him, its center is not the intricate events themselves but rather those who viewed, noted and read out some meaning. Whoever, he says, "pays close attention both to culture and to nature cannot fail to see that the two are bound together." Astronomy is no cultural universal but a product of history that reflects how people live as much as what they see. His imaginative text proceeds thematically. The opening chapter seeks and finds a common ground of process among both ancients and moderns, albeit with disparate mind-set. Then comes a neat summary of the sky seen by the naked eye, with due emphasis on the horizon views that dominate ancient observations up to Alexandrian times. Here we study in particular the comings and goings of Venus, third of the lights, inferior only to sun and moon. Chapter three is a summary of sky myth, "naming the images" in all tongues, brevity forcing some imprecision in detail. Chapter four addresses the astronomy itself: "following the images." Then we examine the systems that explained our life below as signaled by that above, the history of astrology.

Astrology arose long ago and has since died twice. First under the triumph of Christianity within pagan Rome, then after its revival under mediation from Islam, until it lost consensus if not popularity once again after the late Renaissance. The art deserves judgment not on whether its forecasts work but for the participatory opportunities it opens, a means of seeking answers to hard questions even if, as a thriving psychotherapy of gift shop and newspaper, it has become as superficial as fortune cookies. "People first often follow the stars, then come to believe in them," the author says; of course, it is less the stars than their names that they believe in.

Look at astronomy again, the changing lights we can watch, timeless through the paltry span of human history. Begin with Venus brilliant, the morning star, rising just before sunrise to gleam briefly until she is lost in the blaze of a new day. Dawn after dawn she shines longer and longer, until after some eight months she announces dawn for the last time, again close to the horizon. The day after Venus has gone below the horizon, to remain invisible for weeks, in the underworld as the ancients imagined, or simply behind the glowing sun, as a less earth-centered look would confirm. Then she reenters our sky as evening star, brightly decorating twilight, for the same time period, until once more she vanishes, now near earth and in front of the sun for only a matter of days, to return as glorious morning star. So goes the Venus Round endlessly alternating, with minor variations.

But her beat with month or year, never exact, is often close. We have certain clay tablets ascribed to a dynasty of Babylon nearly 40 centuries back, as we hold one folded bark-paper booklet, its number-strewn pages painted 1,000 years ago in red and green. The Venus Tablet of King Ammizaduga and the Mayan text called the Dresden Codex can be read as similar numerical models for the visibility and vanishing of Venus. Both spell out a cycle length near 585 days, the Mayan much closer to the real observable average over a long time, only off by days over centuries, even improved over the long years.

The author makes a good case that both astronomies report short-term data that are adjusted to fit a grander agreement over the long term, fixing the Venus cycle to the solar year and to the lunar month, as our own Roman leap year adjusts the 365-day solar year to fit the genuine quarter-day annual surplus by a shift of one whole day in the four-year cycle. Five entire Venus rounds equal eight solar years within about two days and equal also just 99 lunar months with a discrepancy of less than four days. This is the syncopated beat the old Venus observers and calculators celebrated from Babylon to Copán: an instance of profound unity, with the perceptual precision and professional theory building clearly seen in what we now call science. These were no simple-minded chance discoverers. But the absorbing context, both of religion and polity, shows no such invariant. In Mesopotamia, Venus was personified as Inanna-Ishtar, sensuous and shining, love in its dual aspects. In Mesoamerica the same planet was demonic Kukulcan, ominous yet obscurely parallel to the benign moon herself, by a barely constrained association, as shifting as mere words.

Is the world so plainly ours to invent, to please or even to frighten ourselves? This reader is grateful to Professor Aveni for his account of the current revival of a self-centered view of the natural world. "I would not dare predict—either with lens or horoscopic chart," he says, what will happen. Will one nature be willfully bent to fit shifting and manifold cultures, or will our grasp of cosmic invariants increase? Knowledge and the search for it are not so transient as the cities of river and plain. Venus remains neither demon nor maiden but a rocky globe, whatever the metaphor her scattered rays adorn.

Keeping Up with Computerese

Given the extraordinary growth of computer terminology, it is not enough these days to take pride in avoiding redundant expressions like RAM memory or DOS operating system; no longer a source of linguistic competence to realize that MIPS (Millions of Instructions Per Second) is not the plural of MIP. It is even old hat to have mastered the meaning of RISC (Reduced Instruction Set Computer) microprocessor, that speedy microchip that simplified hardware and shifted many complex operations to the software, thereby earning the alternative name of Relegate Interesting Stuff to the Compiler.

That was just the prelude. Now MPPs (Massively Parallel Processors) are on the horizon. Those of us who absorbed, however imperfectly, three or four generations of computer terminology on the fast trip from Trash 80 (an affectionate name for the old Tandy TRS computer) to Teraflops must meet the challenge. We are going to have to learn to debate the relative merits of SIMD (Single-Instruction, Multiple-Data) and MIMD (Multiple-Instruction, Multiple-Data) machines. Then we'll know about MIMD MPPs.

We will have to learn about MPP topologies (the way the processors are interconnected), including "fat trees," in which processors are grouped into clusters, or even clusters of clusters, and "meshes," in which processors are usually attached in a two-dimensional grid. We will have to conquer the distinction between the "boudoir," with its coupled processor and memory, and the "dance hall," which has processors on one side and memories on the other.

We'll cope. Tera taxonomy, the classification of the massively parallel machines that will soon achieve processing rates of a trillion floating-point operations per second, is, after all, but the latest addition to a dictionary that shows no signs of slowing down. Subduing this lexicon is probably going to be a lifetime task, but given the exuberant nature of the language, it also promises to be a lively one.

Those beginning the job quickly find out that whatever new expression they are encountering will promptly hatch its own linguistic family. For instance, "fuzzy"—initially an adjective in fuzzy logic (the modeling of computer reasoning on the kind of imprecision found in human reasoning), fuzzy representation and fuzzy systems—soon appeared as a verb, "fuzzify," with its nominalization of "fuzzification," leading to the negative form "defuzzify" (to convert to crisp values) and its noun, "defuzzification." The novice confronting this linguistic richness is soon parsing sentences such as "A conventional fuzzy system normalizes and converts its inputs into fuzzy form, executes the rules relevant to the inputs and defuzzifies the resultant output fuzzy sets." And the novice dealing with "compress/decompress" soon discovers that the software that does the job uses two kinds of algorithms, one "lossless," the other "lossy," leading to mastery of the vivid phrase "lossy compression techniques."

Aspirants to the language must also cope with many of its terms in abbreviated forms, usually acronyms (initial letters or parts of the term spoken as a single word, RAM) or initialisms (initial letters or parts of a term pronounced letter by letter, CPU). In line with the ruling ethos of computer terminology—liveliness—these categories are often merged, giving us Troff (Typeset RunOFF, pronounced "tee-roff"), say, or DRAM (Dynamic Random Access Memory, pronounced "dee-ram").

Capitalization in these acronyms comes and goes. Most commonly seen are all capitals (DRAM), but many who find a string of capital letters unsightly, or even distracting, have opted instead to capitalize only the initial letters of longer acronyms. The difference of opinion has led to Fortran (a portmanteau word, from FORmula TRANslation) for some, FORTRAN for others. Inevitably, a linguistic area this rambunctious has given rise to three or even four versions of the same acronym—for instance, Teraflops, teraFLOPS, teraflops and teraFlops (from tera, trillion, plus FLoating-point OPerations per Second).

The flexibility—some might even call it abandon—of computer terminology extends to its cheerful trashing of whatever outdated distinctions remained between nouns and verbs. Consider that classic verb "write," which in computer talk routinely transforms itself into a noun. As the *Communications of the Association for Computing Machinery* (*ACM*) says, "Granularity is the property of memory writes on multiprocessor systems such that independent writes to adjacent aligned data produce consistent results." The fearless interchange of noun and verb frequently leads to what computational linguists call a garden-path sentence guaranteed to be parsed incorrectly on first reading—for example, "The second processor will observe the writes out of order." The popularity of back-to-back nouns in computer talk, with the first one or two nouns used as adjectives ("write latency," "memory write request"), adds to the garden-path effect, because many of the attributive nouns are, of course, ex-verbs. A beginner in the terminology thus acquires patience along with an understanding of such phrases as write latency (it is the time between the memory write request and the storage of the data).

The quickest of looks in dictionaries of computer terms available in print and on-line provides convincing evidence that the terminology is irreverent, upbeat and, above all, changing. So far it seems to be satisfying its expansionist tendencies in the main either by coining words or by wreaking havoc on traditional definitions. It has, however, occasionally preserved the historical meaning of some of its terms, though not for long. "Virtual," for instance, defined by the *Oxford English Dictionary* as "admitting of being called by the name so far as the effect or result is concerned," retained its earlier meaning (in "optics," "virtual focus," "virtual image") even when a few centuries later it became attached to "circuit," "memory" and "disk." Such restraint was probably too much, though, for a language this spirited—and somewhere between memory and disk, and certainly by reality, virtual began to derail into virtuality. Recently it landed on the cover of *Time* magazine for a story on "the dark edges of the computer age," a clear indication that it is time for a virtual replacement. When it arrives, that entertaining reference the *Hacker's Dictionary* is sure to chronicle it. If you are interested in looking up a new word on-line, do as experts in the computer science department advise us—grab it off the net.

ANNE EISENBERG, a professor at Polytechnic University in Brooklyn, is the author of four books on scientific writing.