THE BEST HOPES FOR STOPPING SPAM • MANAGING THE FUTURE

SCIENTIFIC AMERICAN

Monkey Business: The Evolutionary Roots of Economic Behavior

APRIL 2005
WWW.SCIAM.COM

WHAT FLIPS THE MAGNETIC POLES? THE ANSWER LURKS IN THE EARTH'S Geodynamo

Superconductor Breaks the Rules

The Splice of Life: Why Humans Don't Need More Genes

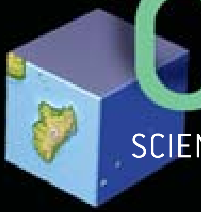
New Painkillers from a Toxic Snail

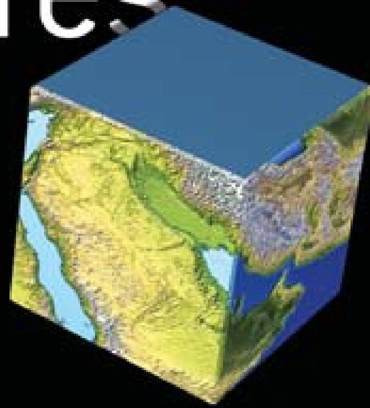
april 2005

contents

SCIENTIFIC AMERICAN Volume 292 Number 4

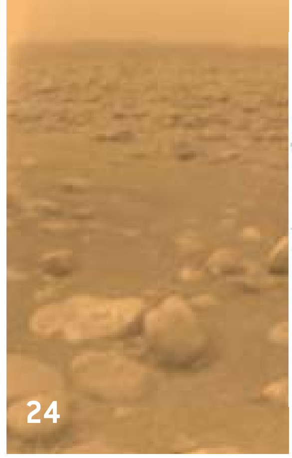
features



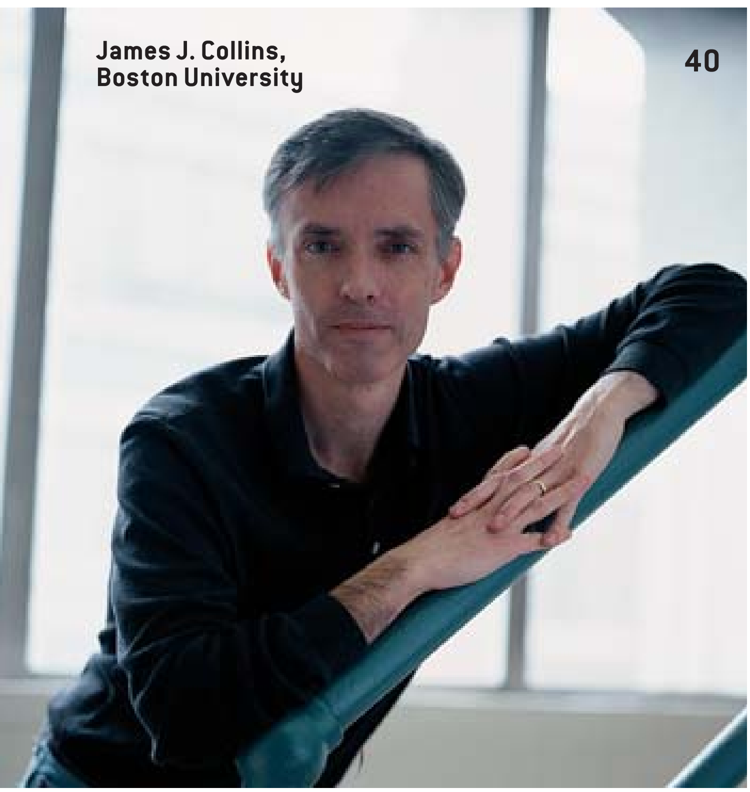


66 Forming flexible policies

departments


24


106


James J. Collins, Boston University

40

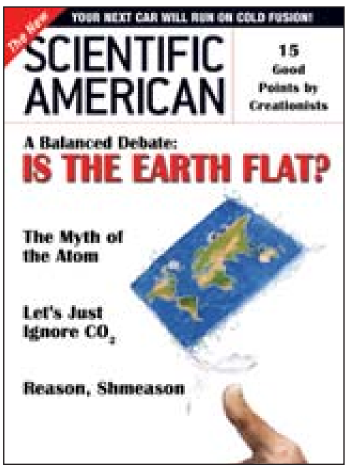
columns

Cover image by Jean-Francois Podevin; photograph at left by Tracy Powell.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 2005 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. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40012504. Canadian BN No. 127387652RT; QST No. Q1015332537. Publication Mail Agreement #40012504. Return undeliverable mail to Scientific American, P.O. Box 819, Stn Main, Markham, ON L3P 8A2. Subscription rates: one year $34.97, Canada $49 USD, International $55 USD. 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; (212) 451-8877; fax: (212) 355-0408 or send e-mail to sacust@sciam.com **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.** Printed in U.S.A.

Okay, We Give Up

There's no easy way to admit this. For years, helpful letter writers told us to stick to science. They pointed out that science and politics don't mix. They said we should be more balanced in our presentation of such issues as creationism, missile defense and global warming. We resisted their advice and pretended not to be stung by the accusations that the magazine should be renamed *Unscientific American,* or *Scientific Unamerican,* or even *Unscientific Unamerican*. But spring is in the air, and all of nature is turning over a new leaf, so there's no better time to say: you were right, and we were wrong.



In retrospect, this magazine's coverage of so-called evolution has been hideously one-sided. For decades, we published articles in every issue that endorsed the ideas of Charles Darwin and his cronies. True, the theory of common descent through natural selection has been called the unifying concept for all of biology and one of the greatest scientific ideas of all time, but that was no excuse to be fanatics about it. Where were the answering articles presenting the powerful case for scientific creationism? Why were we so unwilling to suggest that dinosaurs lived 6,000 years ago or that a cataclysmic flood carved the Grand Canyon? Blame the scientists. They dazzled us with their fancy fossils, their radiocarbon dating and their tens of thousands of peer-reviewed journal articles. As editors, we had no business being persuaded by mountains of evidence.

Moreover, we shamefully mistreated the Intelligent Design (ID) theorists by lumping them in with creationists. Creationists believe that God designed all life, and that's a somewhat religious idea. But ID theorists think that at unspecified times some unnamed superpowerful entity designed life, or maybe just some species, or maybe just some of the stuff in cells. That's what makes ID a superior scientific theory: *it doesn't get bogged down in details*.

Good journalism values balance above all else. We owe it to our readers to present everybody's ideas equally and not to ignore or discredit theories simply because they lack scientifically credible arguments or facts. Nor should we succumb to the easy mistake of thinking that scientists understand their fields better than, say, U.S. senators or best-selling novelists do. Indeed, if politicians or special-interest groups say things that seem untrue or misleading, our duty as journalists is to quote them without comment or contradiction. To do otherwise would be elitist and therefore wrong. In that spirit, we will end the practice of expressing our own views in this space: an editorial page is no place for opinions.

Get ready for a new *Scientific American*. No more discussions of how science should inform policy. If the government commits blindly to building an anti-ICBM defense system that can't work as promised, that will waste tens of billions of taxpayers' dollars and imperil national security, you won't hear about it from us. If studies suggest that the administration's antipollution measures would actually increase the dangerous particulates that people breathe during the next two decades, that's not our concern. No more discussions of how policies affect science either—so what if the budget for the National Science Foundation is slashed? This magazine will be dedicated purely to science, *fair and balanced* science, and not just the science that scientists say is science. And it will start on April Fools' Day.

MATT COLLINS

THE EDITORS editors@sciam.com

How to Contact Us

EDITORIAL

For Letters to the Editors:
Letters to the Editors
Scientific American
415 Madison Ave.
New York, NY 10017-1111
or
editors@sciam.com
Please include your name and mailing address, and cite the article and the issue in which it appeared. Letters may be edited for length and clarity. We regret that we cannot answer all correspondence.

For general inquiries:
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-754-0550
fax: 212-755-1976
or
editors@sciam.com

SUBSCRIPTIONS

For new subscriptions, renewals, gifts, payments, and changes of address:
U.S. and Canada
800-333-1199
Outside North America
515-248-7684
or
www.sciam.com
or
Scientific American
Box 3187
Harlan, IA 51537

REPRINTS

To order reprints of articles:
Reprint Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8877
fax: 212-355-0408
reprints@sciam.com

PERMISSIONS

For permission to copy or reuse material from *SA*:
www.sciam.com/permissions
or
212-451-8546 for procedures
or
Permissions Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
Please allow three to six weeks for processing.

ADVERTISING

www.sciam.com has electronic contact information for sales representatives of Scientific American in all regions of the U.S. and in other countries.

New York
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8893
fax: 212-754-1138

Los Angeles
310-234-2699
fax: 310-234-2670

San Francisco
415-403-9030
fax: 415-403-9033

Detroit
Karen Teegarden & Associates
248-642-1773
fax: 248-642-6138

Midwest
Derr Media Group
847-615-1921
fax: 847-735-1457

Southeast
Publicitas North America, Inc.
404-262-9218
fax: 404-262-3746

Southwest
Publicitas North America, Inc.
972-386-6186
fax: 972-233-9819

Direct Response
Special Aditions Advertising, LLC
914-461-3269
fax: 914-461-3433

Belgium
Publicitas Media S.A.
+32-(0)2-639-8420
fax: +32-(0)2-639-8430

Canada
Derr Media Group
847-615-1921
fax: 847-735-1457

France and Switzerland
PEM-PEMA
+33-1-46-37-2117
fax: +33-1-47-38-6329

Germany
Publicitas Germany GmbH
+49-211-862-092-0
fax: +49-211-862-092-21

Hong Kong
Hutton Media Limited
+852-2528-9135
fax: +852-2528-9281

India
Convergence Media
+91-22-2414-4808
fax: +91-22-2414-5594

Japan
Pacific Business, Inc.
+813-3661-6138
fax: +813-3661-6139

Korea
Biscom, Inc.
+822-739-7840
fax: +822-732-3662

Middle East
Peter Smith Media & Marketing
+44-140-484-1321
fax: +44-140-484-1320

Scandinavia and Finland
Publicitas Nordic AB
+46-8-442-7050
fax: +46-8-442-7059

U.K.
The Powers Turner Group
+44-207-592-8331
fax: +44-207-630-9922

On the Web

WWW.SCIAM.COM

UPDATED EVERY WEEKDAY

Visit www.sciam.com/ontheweb to find these recent additions to the site:

Secret of the Venus Flytrap Revealed

At just a tenth of a second, the snapping action that a Venus flytrap uses to capture its prey is one of the fastest movements in the plant kingdom. Scientists have long wondered how the plant manages such a feat without muscles or nerves. The answer, it appears, is by shape shifting.

Chunk of Universe's Missing Matter Found

In recent years, astronomers have found themselves faced with a nagging inventory problem. Received wisdom holds that dark matter and dark energy make up 95 percent of the universe and that ordinary matter, or baryons—the subatomic particles that form planets, stars and the like—account for the remainder. The problem is, the luminous matter detected with the aid of optical telescopes has amounted to a mere 10 percent of the expected ordinary matter, and the baryons inferred by other means bring that total to only 50 percent. New findings are helping to bridge this gap between prediction and observation.

Study Identifies Trade-off between Motherhood and Longevity

Motherhood is a difficult job. A recent study found that, historically, taking on the role early in life was linked to shorter life span. The results suggest that natural selection may have sacrificed a woman's longevity for reproductive success.

Ask the Experts

How does anesthesia work?
Bill Perkins, professor of anesthesiology at the Mayo Clinic College of Medicine in Rochester, Minn., explains.

E-mail Newsletters & Alerts

Stay informed with Scientific American.com's Weekly Review, delivering the latest science and technology news every Tuesday.

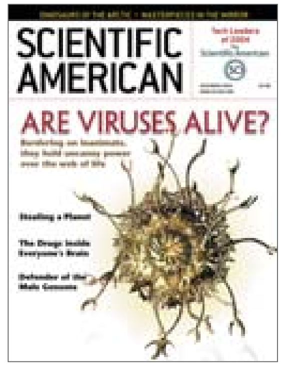
Sign up today.

Visit www.sciam.com/newsletter

COURTESY OF FORTERRE AND MAHADEVAN

Letters

EDITORS@SCIAM.COM



READERS WHO tucked the December 2004 issue into their travel bags for the holidays could take an armchair tour of a wide assortment of topics and times. They could hike through new evidence about the 19th-century Anglo-French controversy over whose citizen-scientist discovered Neptune first; reflect on the cover story's very 21st-century question, "Are viruses alive?"; and take in culture by exploring whether or not 15th-century artists employed optical projections. Those preferring more exotic climes and times, say, the Cretaceous period, could track Arctic Alaskan dinosaurs.

December also featured the "Scientific American 50," which prompted Brian Miller of Woodland Hills, Calif., to write: "I sincerely want to thank all the outstanding minds who bring us these innovations. And thanks also to *Scientific American* for giving them the credit they so seldom receive."

OPTICAL REJECTION

Regarding "Optics and Realism in Renaissance Art," author David G. Stork did not read my book with care. I said that to see an optical projection is to use it. When one knows how simple cameras are—they can be just a small concave mirror—and that the invention of photography is the invention of a chemical process, not one of optics, one can see that the arrival of chiaroscuro and perspective in Western art has something to do with optics.

My book is simply relooking at the history of European painting assuming they knew about optical projections. The reason all this is interesting today is that with the advent of digital photography the chemicals have been removed from photography. We are in a new era.

The camera can be used as it was in the past, with collage techniques to make new pictorial spaces. This offers exciting new possibilities that have connections with image making from the past. Chemical photography was the end of something; now we have an exciting new beginning. Its relationship with reality will be the same as drawing and painting. The hand is back in the camera, and photography's claims to veracity are now foggy.

David Hockney
Los Angeles

STORK REPLIES: I am grateful if Hockney is retreating from the claim that early Renaissance masters directly traced projected images—which he and Charles Falco have claimed explicitly and presented to millions of viewers. As other experts and I have shown, this claim fails on a number of grounds.

REEFER MADNESS

"Marijuana Research" [SA Perspectives], which described cannabinoids' potential benefits, was excellent. I can't say that your publication has always taken such a position. A March 1936 *Scientific American* news item, "Marihuana Menaces Youth," stated: "Marihuana produces a wide variety of symptoms in the user, including hilarity, swooning and sexual excitement. Combined with intoxicants, it often makes the smoker vicious, with a desire to fight and kill."

I'm glad that you've changed your position over the past 70 years. I hope Congress will follow suit.

Jp Doiron
Edmonton, Alberta

FINDERS KEEPERS

As the scholar you accurately credited as discoverer of the prime scientific justification for the startling conclusion of "The Case of the Pilfered Planet," by William Sheehan, Nicholas Kollerstrom and Craig B. Waff, I have some comments to submit regarding this article.

Adams's final September 2, 1846, extrapolation solution for Neptune's position was not a minor afterthought. It was quite definite (315 degrees, 20 minutes longitude), quite wrong (off by over 10 degrees), and thus quite crucial in causing Britain to lose the race.

One of our key discoveries in the long-hidden, now recovered Neptune file is Adams's October 15, 1846, letter of explanation to Astronomer Royal George Airy declaring regret for so hugely misdirecting the British search. This final position was linearly extrapolated from two perturbational-math solutions for slightly different mean distances, an intelligent approach but time-consuming, therefore closely related to later historians' misperception that Adams was psychologically paralyzed.

Dennis Rawlins
DIO, The International Journal of Scientific History
Baltimore, Md.

I disagree that Adams does not deserve equal credit for Neptune's discovery. He did predict the planet's existence before French mathematician Urbain Jean Joseph Le Verrier, and he communicated it to British Astronomer Royal George Biddle Airy and astronomer James Challis. He could be blamed for not pushing his finding forcefully enough but not for the two astronomers' failure to act on it.

Also, Adams may have referred to the concept of a hypothetical mass tugging on Uranus as an "abstraction," but the implication was nonetheless clear that it referred to a new planet.

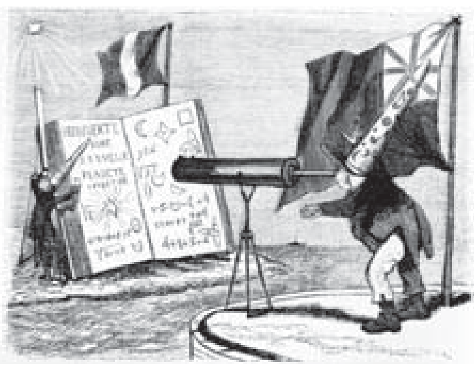
I also do not agree that "luck" played a role in both Le Verrier's and Adams's ability to make reasonably accurate predictions based on orbits that were, in hindsight, very different from the real orbit. It can easily be shown that there exists quite a large possible set of orbits and an associated massive object that could have approximated the gravitational force vector of Neptune on Uranus in the period between 1800 and 1850—during which it was thought the most reliable observations were made. In most of these cases, the apparent position of the hypothetical planet would not have deviated far from Neptune's actual position as seen from Earth at the time, ensuring a close prediction.

Cornelis F. du Toit
NASA Goddard Space Flight Center

SHEEHAN, KOLLERSTROM AND WAFF REPLY: In his September 2, 1846, letter to Airy (available in the "Account" that Airy presented to the Royal Astronomical Society in November 1846), Adams not only provided extensive details of his second perturbation solution but also briefly discussed the extrapolation estimate mentioned by Rawlins. That year Adams explained to Airy that in his previous letter "I inferred that the mean distance used in my first hypothesis [sent to Airy in October 1845] must be greatly diminished, but I rather hastily concluded that the change in the mean Longitude deduced would be nearly proportional to the change in the assumed mean distance." There is nothing in the latter letter that can be interpreted as Adams expressing regret "for so hugely misdirecting the British." Quite the contrary, no evidence has surfaced that Challis was aware of the extrapolation estimate or that he immediately and significantly changed his search in response to it.

Regarding du Toit's remarks, the article stated that Adams's failure to communicate his fall 1845 prediction forcefully to Challis and Airy (particularly by not including his methodology) did contribute to their separate decisions not to initiate a planet search before late June 1846. Challis was skeptical that analysis could predict a position with sufficient accuracy to limit the search to a reasonably small area of sky. Also, Airy had his hang-up regarding the radius vector.

It was only the nearly identical (to Airy at least) predicted positions of Adams and Le Verrier, plus the latter's discussion of his methodology in his June 1 paper, that finally set Airy and Challis into action, coincidentally at a time (near opposition for the hypothetical planet) that was favorable for sighting it if it actually existed.



CARTOON POKES FUN at Britain's co-claim to Neptune's discovery in the November 7, 1846, *L'Illustration*.

ARCHIVES CHARMET

GENERAL AVIATION RULES

"Crowded Skies," by Mark Fischetti [Working Knowledge], on air traffic control was accurate as far as airliners are concerned, but general aviation aircraft have options you didn't mention.

First, general aviation aircraft, even operating under visual flight rules, can enter controlled airspace provided they are identified on radar by the controller. Second, if the pilot has a license to fly using instruments, she can file a flight plan under instrument flight rules and then receive the same treatment as the big airplanes, with radar following all the way from takeoff to landing. The actual possibilities are more complicated, but this gives the general idea.

Robert West
Madison, Wis.

FISCHETTI REPLIES: We oversimplified the description of the airspace where general aviation craft can fly, implying they can navigate only outside that governed by air traffic controllers while using only visual flight rules.

ERRATA The *Albertosaurus* pictured in "The Dinosaurs of Arctic Alaska" on pages 88 and 89 is shown with three digits on each of its upper limbs, rather than two.

On page 57 of "The Scientific American 50," in the description of Nantero, it was incorrectly stated that SRAMs retain data when the power is off. They do not.

In "The Case of the Pilfered Planet," Isaac Newton's early home should have been given as Lincolnshire, England.

50, 100 & 150 Years Ago
FROM SCIENTIFIC AMERICAN

Nobel Viruses ▪ Noble Gases ▪ King of Fertilizers

APRIL 1955

VIRUS REPLICATION—"A new view of the nature of viruses is emerging. They used to be thought of solely as foreign intruders—strangers to the cells they invade and parasitize. But recent findings, including the discovery of host-induced modifications of viruses, emphasize more and more the similarity of viruses to hereditary units such as genes. Indeed, some viruses are being considered as bits of heredity in search of a chromosome. —Salvador E. Luria" [*Editors' note: Luria shared the Nobel Prize in Physiology or Medicine in 1969 for his work on the replication and interaction of viruses.*]

POLIO VACCINE—"We shall soon learn the results of last year's extensive field test of the vaccine against poliomyelitis. Whatever the analysis of that test shows, the type of vaccine that is being tested will continue to be an issue among virologists, because an immunological principle is under test as well as a vaccine. The vaccine in question is made of a 'killed' virus, that is, a virus rendered noninfectious by treatment with formaldehyde. Many virologists believe such a vaccine can never be as effective as one containing live virus. I share the view that a killed-virus vaccine not only avoids the hazards of live virus but, if properly prepared and used, may be just as effective in producing immunity. —Jonas E. Salk"

CULTURAL CUES—"Unfortunately, many of the most important standards of acceptable behavior in different cultures are elusive: they are intangible, undefined and unwritten. In the interest of intercultural understanding, various U.S. Government agencies have hired anthropologists from time to time as technical experts. For instance, in Latin America, where touching is more common and the basic units of space seem to be smaller, North Americans are disturbed by how close the Latin Americans stand when they converse. The Latin Americans, for their part, complain that people in the U.S. are distant and cold—*retraídos* (withdrawing and uncommunicative) [*see illustration*]."

A CULTURAL PERCEPTION scrutinized by anthropologists, 1955

APRIL 1905

RAMSAY'S WORK—"Some time ago Sir William Ramsay communicated to the Royal Society estimates of the amounts of krypton and of xenon in atmospheric air, and since then he has been doing the same for neon and helium. After a series of delicate investigations, he arrived at the conclusion that there are in gaseous air 86 parts by weight of neon in a thousand million, while of helium there are 56 parts by weight, and 400 by volume, in ten thousand million. Such minute amounts seem almost incalculably small, but corroborative tests had been applied, which indicated that the estimates could not be far from accurate."

APRIL 1855

TOOTH DECAY—"There is no good reason why the teeth of our people are so subject to early decay, in comparison with the teeth of the people of some other countries. It is generally allowed that there is work for five times the number of dentists in the United States than there is in Britain; and that, while bad teeth is the exception there, it is the rule here. We believe that our people take more pains with their teeth by washing than the natives of Ireland, and yet the Irish have far better teeth. Acetic acid cannot be the cause of this early decay of teeth among us; indeed, we know it is not. It is our opinion that if more coarse *hard* biscuit were eaten in early life, to properly exercise the teeth, they would be less liable to early decay."

GUANO—"It is fifteen years since guano was first used in Europe, when only a few tons were used, by way of experiment; since then its consumption has increased up to nearly 100,000 tuns per annum. In our own country the demands for it were so numerous last year that they could not be supplied. From the great and insatiable demand for this manure, it has been subject to the most glaring adulterations by mixing it with loam of the same color as the guano itself. Farmers should therefore purchase it of respectable and well-known dealers. The best quality of this fertilizer is the Peruvian."

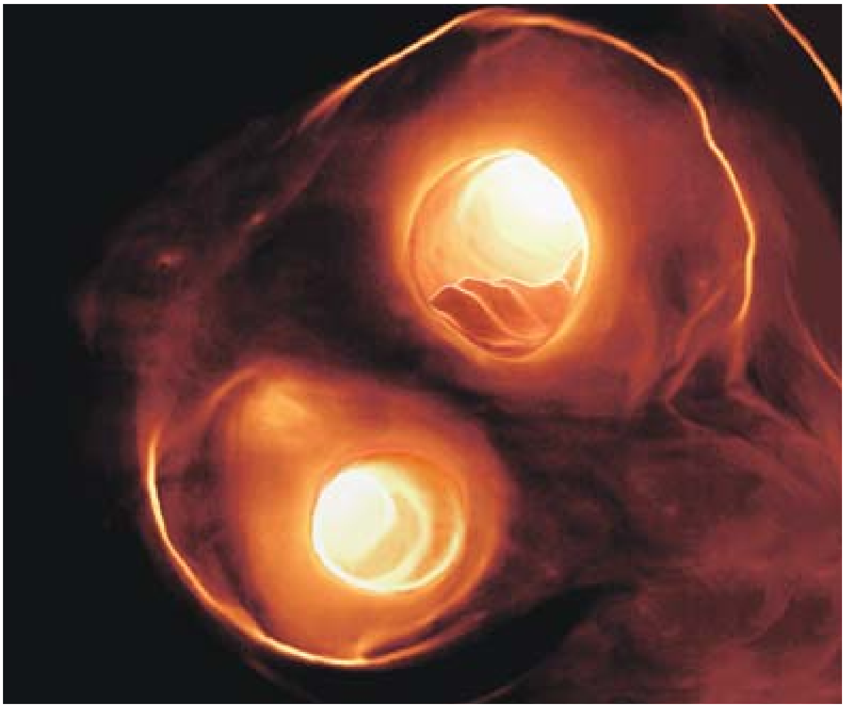
news SCAN

HEART DISEASE

Reactive Reasoning

IS AN INFLAMMATION PROTEIN THE NEXT CHOLESTEROL? **BY DIANE MARTINDALE**

Calming inflammation in the body might be just as important for fighting heart disease as lowering cholesterol, according to two studies released in January. Some experts see the findings as evidence to more aggressively monitor and perhaps even treat inflammation in patients. But others are not yet convinced that doing so would extend more lives.



FATTY DEPOSITS called plaques (*rippled area in orange*) build up in blood vessels as revealed in this CT scan of the carotid artery where it splits into its two main branches. The plaques may trigger the production of C-reactive protein, a possible factor in heart disease.

A key mediator in atherosclerosis, inflammation damages the lining of the artery walls and contributes to the formation and rupture of fatty plaques. Starting in 1997, Paul M. Ridker, a cardiologist at Brigham and Women's Hospital, began finding a relation between an inflammation compound called C-reactive protein (CRP) and heart disease. For instance, CRP levels of more than about two milligrams per liter of blood almost tripled the risk for heart attack and doubled the risk for stroke.

In two separate studies published in the January 6 *New England Journal of Medicine,* Ridker and Steven E. Nissen of the Cleveland Clinic looked at nearly 4,300 patients with severe heart disease who were taking moderate or high doses of statin drugs to lower their cholesterol. They wondered why some patients seemed to be doing better than others despite having reached the same low levels of LDL, the bad cholesterol.

Ridker confirmed his earlier findings, and Nissen found a link between lower CRP levels and a regression of atherosclerosis (the plaques got smaller). More important, the reductions in CRP had a beneficial effect independent of lowering LDL, suggesting that statins lessen both cholesterol and CRP.

GJLP Science Photo Library

The reports indicate that cutting CRP is at least as critical as lowering cholesterol, Ridker says, and "support the concept that CRP is not only a useful clinical marker of inflammation but also a player in heart disease." Nissen agrees: "In the future, we will attack CRP with the same vigor that we attack cholesterol." Ridker suspects that healthy individuals who have normal cholesterol (130 milligrams per deciliter of blood) but high CRP levels might also benefit from taking statins and has started a new trial of 15,000 people to explore the possibility.

Despite the strong evidence, other experts warn that more work is needed to prove that CRP directly causes atherosclerosis or that people should take statins to control CRP. "These studies tell me that drugs that lower both LDL and CRP may have a greater effect than therapies that lower just LDL," says David S. Siscovick, co-director of the cardiovascular health research unit at the University of Washington School of Medicine. "But this is not going to change the way I treat my patients." That's because statins may not reduce CRP directly. Instead the drugs may interfere with the inflammatory pathway in the body, and less inflammation may reduce cardiovascular risk. If so, then CRP is at most a marker, not a cause, of heart disease.

In fact, the mechanisms that raise CRP levels are not entirely clear. Infections and chronic diseases such as rheumatoid arthritis as well as obesity, smoking, high blood pressure and diabetes all boost CRP levels. And when a person loses weight, quits smoking, and controls diabetes and blood pressure, their CRP levels fall, indicating that CRP is a marker of these other problems associated with inflammation.

What is more, some experts doubt the usefulness of CRP for patient screening in the clinic. "It doesn't help me determine who is at risk—who I should and shouldn't treat," remarks Donald Lloyd-Jones, a cardiologist at Northwestern University. According to Lloyd-Jones, who is also an epidemiologist and has explored risk factors in heart disease, there has been too much emphasis on the fact that CRP can give a statistically significant relative risk but not enough focus on whether it actually adds to our ability to discriminate risk. For instance, using established risk factors such as cholesterol level and obesity, physicians can tell who will get heart disease from those who will not 80 percent of the time. When CRP is added to the mix, Lloyd-Jones notes, that number goes up to 81 percent. "It turns out that this is statistically significant, but it doesn't help me as a doctor," he says.

Such a viewpoint is certain to disappoint the statin drugmakers, which sponsored the latest studies. "This train of CRP has so much momentum," Lloyd-Jones adds. "But if we look at it closely, it's not what it's cracked up to be."

Diane Martindale is based in Toronto.

TROUBLESOME FATS

Some researchers suggest that C-reactive protein (CRP) levels rise during the development of heart disease because of the inflammation triggered by the fatty plaques lining the coronary arteries. But others are not convinced that the tiny deposits can set off high CRP production. Instead they favor the idea that fat tissue, especially around the waist, acts as an inflammatory organ and that macrophages, which invade fat tissue, send out signals to the liver to make more CRP. If indeed fat cells stimulate CRP production more robustly than the plaques in arteries do, then high CRP levels may indicate risk factors related to obesity rather than heart disease directly.

PLANETARY SCIENCE

Strange New World

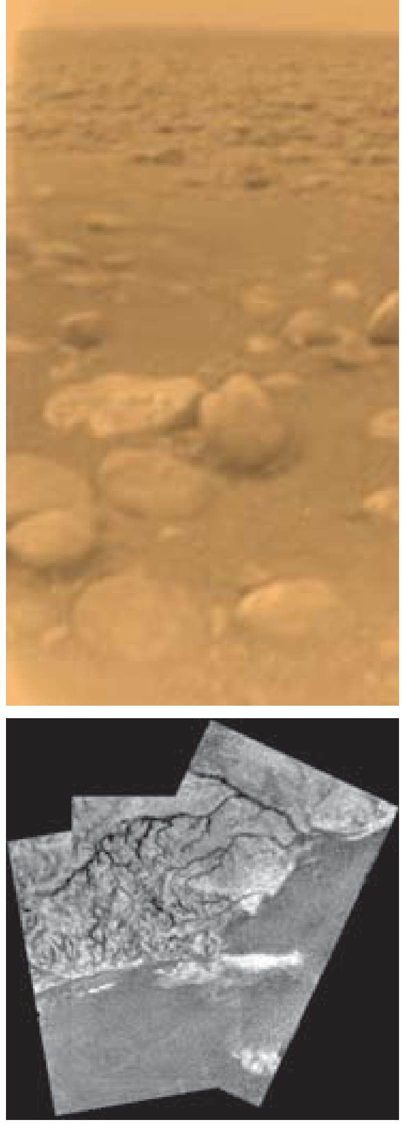
PIERCING THE HAZE, HUYGENS GETS A VIEW OF TITAN'S SURFACE **BY MARK ALPERT**

On January 14 a saucer-shaped spacecraft weighing 320 kilograms made the "splat" heard round the solar system. The successful landing of the Huygens probe on Titan, Saturn's large and mysterious satellite, delighted planetary scientists, who thrilled at the probe's images of icy ridges and dark, riverlike channels. In the following weeks, though, the euphoria turned to head-scratching as researchers struggled to decipher the data collected by the probe. Although the four-hour-long mission provided the first close look at Titan's surface and atmosphere, Huygens raised at least as many questions as it answered.

Despite the moon's extreme cold—its surface temperature is –180 degrees Celsius—Titan is similar to Earth in many ways. Like our planet, Titan has a thick atmosphere composed mainly of nitrogen. Another significant constituent is methane, which condenses at low temperatures and appears to play the same meteorological role on Titan as water does on Earth. Scientists had long speculated that Titan might have clouds of methane and lakes or seas of liquid hydrocarbons. To test this hypothesis, the European Space Agency built the Huygens probe, which for seven years journeyed to Saturn attached to NASA's Cassini orbiter. The two craft separated as they approached Titan, and Cassini relayed Huygens's data to Earth after the probe parachuted through the satellite's haze, which had obscured previous attempts to observe the surface.

One of the mission's first surprises was that this haze extended so far down. Titan's main haze layer is at least 200 kilometers thick; the probe's view of the surface did not clear until Huygens was less than 20 kilometers above the ground. At that same altitude the onboard gas chromatograph and mass spectrometer detected an abrupt jump in the amount of methane in the atmosphere. Together the findings suggest the presence of methane clouds, with the droplets possibly condensing around the haze particles. (These hydrocarbon solids are believed to come from the breakdown of methane in Titan's upper atmosphere caused by the sun's ultraviolet radiation.)

Evidence of methane rainfall came from Huygens's images of the surface, which showed sinuous, branching channels extending from relatively bright highlands to a tarry plain. The contrasts in brightness suggest that precipitation may have washed dark hydrocarbon deposits off the icy highlands and into the channels. But some of the channels are short and stubby, leading scientists to speculate that liquid methane may also be flowing from underground springs.



IMAGES OF TITAN: Icy pebbles litter the satellite's surface (*top*); the rounding of the pebbles is evidence of erosion, possibly caused by liquid flow. During its descent, the Huygens probe saw an icy ridge about 100 meters high and a network of riverlike channels (*bottom*).

ESA/NASA/JPL/UNIVERSITY OF ARIZONA (*top and bottom*)

The probe's landing, though, offered the strongest indications of liquid methane on Titan. As Huygens hit the surface at five meters per second, a device called a penetrometer—essentially a spring-loaded stick on the underside of the craft—measured the force of impact and found the ground's resistance to be like that of wet sand with a brittle crust. Three minutes after landing, Huygens detected a 30 percent rise in the abundance of methane; the heat of the probe's instruments had apparently vaporized the liquid hydrocarbon in the top few centimeters of Titanian soil.

Researchers will spend the next several months analyzing the data and devising theories to explain the observations. "Titan is living up to our expectations," says Jonathan I. Lunine, a University of Arizona scientist on the Cassini-Huygens team. "It's as interesting as we'd hoped it would be." Tobias C. Owen, an atmospheric scientist at the University of Hawaii's Institute for Astronomy, warns that theorists must be careful about generalizing the results from a single landing site; it may not be possible, for example, to estimate the frequency of methane rainfall on Titan or the total amount of liquid methane on the surface. But continuing observations by the Cassini orbiter, which scanned Titan's surface near the landing site in a flyby in February, may help investigators make sense of Huygens's remarkable findings.

BLOWING IN THE WIND

One glitch marred the Huygens mission to Titan: the operations team failed to turn on a radio receiver on the Cassini orbiter that was supposed to gauge wind speeds on the Saturnian moon by measuring Doppler shifts in the frequency of an ultrastable signal sent by the probe during its descent. But radio telescopes on Earth managed to catch the feeble signal, allowing scientists to recover some of the wind-speed data. The buffeting of the probe indicated very strong winds (430 kilometers per hour, about the same as the fastest jet-stream winds on Earth) at 120 kilometers above the surface and large variations in wind speed, possibly caused by vertical wind shear, at altitudes between 60 and 80 kilometers.

HEALTH

Leafy Letdown

EATING VEGETABLES SEEMS TO DO LITTLE IN WARDING OFF CANCER **BY JR MINKEL**

Munch on fruits and vegetables, health officials implore—they will reduce the risk of chronic ills such as heart disease, high blood pressure and cancer. But large, long-term studies have mostly failed to corroborate initial signs of the cancer-fighting powers of produce. Recent studies suggest that any protective effect, especially for breast cancer, is weak or confined to subgroups of people.

The 1970s and 1980s saw hundreds of reports documenting dips in cancer rates of relatively voracious veggie eaters. In most of these studies, researchers asked people newly diagnosed with cancer to reconstruct their recent diets and compared the results with those of healthy volunteers. Such "case control" studies can be misleading, however. People in the cancer group may overestimate their consumption of foods they perceive as harmful, and volunteers tend to be more health-conscious than the general population. Nevertheless, the findings persuaded the National Cancer Institute to launch the national 5-a-Day Program in 1991 to take advantage of the apparent effect, encouraging people to eat five to nine servings of fruits and vegetables daily.

Stronger evidence can emerge from clinical trials and prospective, or cohort, studies, in which researchers interview large groups of healthy people about their diet and follow them over time. Results have begun coming in during the past five years. In a surprise, the findings failed to turn up protective effects against colon, lung and breast cancer. But negative results may simply mean the investigators were not looking in the right places. After all, diet is notoriously variable and difficult to quantify. (Try to catalogue what you ate in the past month.) And at least in the case of breast cancer, researchers might have misclassified fruit and vegetable consumption by using different questionnaires for different groups.

Work published last November seems to strengthen the case against a general protective effect. A questionnaire-based study that combined cohorts of American doctors and nurses found no change in cancer incidence based on fruit and vegetable consumption, although it did turn up a 25 percent decrease in cardiovascular disease among the heaviest consumers. The outcomes, reported by Harvard School of Public Health scientists in the *Journal of the National Cancer Institute,* suggest that any protective effect for cancer must be small, says Lawrence H. Kushi, associate director for etiology and prevention research at Kaiser Permanente in Oakland, Calif.

That conclusion gained further support in relation to breast cancer in January, when researchers published results from the European Prospective Investigation into Cancer and Nutrition (EPIC) study, the biggest of its kind yet. A team from the University Medical Center of Utrecht in the Netherlands examined five years' worth of data from 286,000 women aged 25 to 70. Participants spanned eight countries from northern to southern Europe, with foods ranging from sauerkraut to eggplant parmesan. The researchers, publishing in the *Journal of the American Medical Association,* found no difference in breast cancer rates based on total consumption of fruits and vegetables or on intake of specific vegetable groups, such as leafy greens or roots.

A protective effect might still lurk in the data, however. For one thing, none of the studies has probed childhood diet. In the

DAVID R. FRAZIER Photo Researchers, Inc.

STILL A HEALTHY CHOICE, THOUGH: Fruits and vegetables offer little protection against cancer.

THE FOOD OR THE GENES?

Dietary cancer protection, if it exists, may depend more on metabolism than on the amount of vegetables eaten. In one study involving nurses, the same participants who showed no change in cancer based on vegetable intake differed 25 to 35 percent in their breast cancer rates. Turns out that those with the most carotenoids (including beta-carotene, found in carrots) in their blood had the least incidence of breast cancer, researchers report in the January 15 *American Journal of Epidemiology.* That would seem to indicate that eating more carrots lowers the risk of breast cancer. "But the amount of carotenoids in the blood doesn't correlate perfectly with the amount of food [eaten]," notes Lawrence H. Kushi, a cancer expert at Kaiser Permanente in Oakland, Calif. Metabolic differences may make some women better at harvesting protective nutrients from their food.

case of breast cancer, young girls may pass through a critical period during which fruits and vegetables could block the disease's early stages. Another factor could be genetic differences in anticancer enzymes, such as those that scavenge harmful free radicals, highly reactive compounds that damage DNA. Some studies have found differences in cancer rates in people who have specific gene variants and consume particular vegetables, notes Christine B. Ambrosone, an epidemiologist at Roswell Park Cancer Institute in Buffalo, N.Y. "I have a little statue of broccoli on my desk," she says. "I don't think these cohort findings are the final word."

Despite the uncertainty, experts agree the five-a-day recommendation still holds as part of overall healthy eating habits. So don't ditch the fruit salad and crudités just yet.

JR Minkel is a frequent contributor.

IMAGING

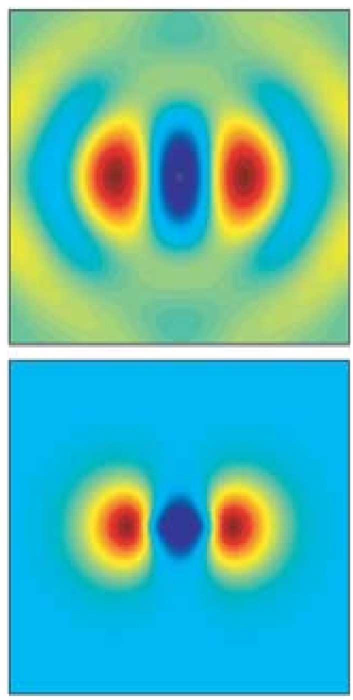
CT Scan for Molecules

PRODUCING 3-D IMAGES OF ELECTRON ORBITALS BY GRAHAM P. COLLINS

In the quantum world, objects are described by wave functions. Electrons around a molecule, for example, exist in wavelike orbitals, smeared-out shapes that determine properties such as the electrons' energy and the propensity for the molecule to undergo various chemical reactions. But orbitals are slippery critters that, because of Heisenberg's uncertainty principle, defy routine efforts to image them completely and accurately. Now, however, researchers at Canada's National Research Council in Ottawa have produced a three-dimensional scan of the outermost electron orbital around a nitrogen molecule. The "shutter speed" of the imaging method is fast enough that scans might one day be taken of molecules caught in the middle of a chemical reaction.

The group, led by Paul B. Corkum and David M. Villeneuve, uses a laser pulse lasting just 30 femtoseconds (3×10^{-14} second). During the course of the laser pulse, the electric field of the light wave oscillates about a dozen times. Each oscillation drives the outermost electron of the nitrogen molecule away from the molecule and back again.

Although it might seem that the team relies on a laser to "light up" the electron, it is actually the electron on its way back toward the molecule that acts as the imaging beam. More precisely, the laser's field drives a small proportion of the electron's wave function away and back. Think of it as the electron being in two places at once; mostly it is still in place in its original orbital around the nitrogen, but partly it is being ripped away.



ELECTRON ORBITAL of a nitrogen molecule as imaged (*top*) agrees quite well with the orbital as calculated from theoretical models (*bottom*). Each image is 0.6 nanometer by 0.6 nanometer in size. Colors represent the amplitude of the quantum wave function—the electron is most likely to be found at the red and dark blue locations.

DAVID M. VILLENEUVE

The sharp acceleration turns the traveling electron wave into a plane wave, like a nice regular pulse of an electron beam with an extremely short wavelength—exactly the kind of beam useful for imaging. When the plane wave returns and crosses the molecule, it produces an interference pattern with the stationary part of the electron wave function, like two trains of water waves crossing and forming a checkerboard disturbance.

To complete the imaging, that interference pattern must be detected. As the plane wave travels along, the pattern oscillates rapidly, causing it to emit ultraviolet radiation that the researchers observe. Information about the shadow of the electron orbital as seen by the traveling electron wave is imprinted on the ultraviolet emission. Producing a three-dimensional image requires repeating the process at different angles, like a hospital CT scanner. The angles are set by aligning all the nitrogen molecules in the sample with a somewhat weaker laser pulse a few picoseconds (10^{-12} second) before the imaging pulse arrives.

The result of the imaging agrees quite well with the shape of the electron orbital computed theoretically. "I was very excited when I saw the experimentally obtained images of molecular orbitals for the first time," says Ferenc Krausz of the Max Planck Institute of Quantum Optics near Munich. "The technique has great potential." In late 2003 Krausz's group demonstrated another kind of imaging using 250-attosecond (2.5×10^{-16} second) pulses of extreme ultraviolet light, the shortest light pulses ever produced. The two methods are complementary—Krausz's involving the dynamics of inner electrons, Corkum and Villeneuve's working on the outermost electrons.

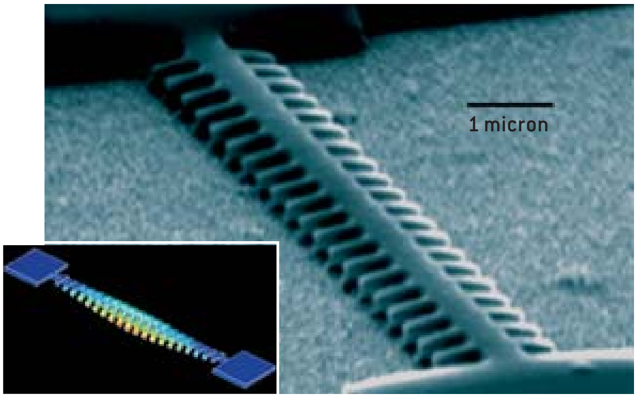
Of great interest will be the application of the technique to more complicated molecules and to molecules caught in the process of undergoing a chemical reaction. Villeneuve says he is considering trifluoromethyl iodide, which can be broken up by pulses from the group's laser. "Then we could follow the dissociation," he says, "and measure how the atoms move."

COMPUTING

Qubit Twist

BENDING NANOTUBES AS MECHANICAL QUANTUM BITS **BY CHARLES Q. CHOI**

Before the advent of electricity, the first computers were mechanical, with the Difference Engine invented by Charles Babbage tackling logarithms and trigonometry 150 years ago. Now advanced quantum computers might go back to mechanical roots, using rows of nanometer-scale bars as moving parts.



SILICON SWING: This multielement bar, consisting of 50 billion silicon atoms, could flex when driven by an oscillating force, according to a computer model (*inset*). Such bars might be the basis for mechanical quantum bits.

The bizarre laws of quantum physics suggest that items the size of molecules and smaller can exist in two or more places or states at the same time. An observation or some other action forces them to collapse out of this "superposition," leading to just one outcome. In theory, because quantum bits, or "qubits," can exist in both an on and off state simultaneously, a quantum computer with just 300 qubits can run more calculations in an instant than there are atoms in the universe.

Existing methods to create qubits rely on trapping atoms with lasers or manipulating nuclear spins in semiconductor crystals, among other approaches. These techniques, however, are highly delicate, and the slightest disturbance can disrupt the superposition of qubits prematurely. At best, researchers have managed to "entangle," or connect up, only a few qubits to form simple logic operations.

A more robust alternative might be mechanically based qubits. Imagine a tape measure extended a few centimeters, explains theoretical physicist Franco Nori of the University of Michigan at Ann Arbor and the Frontier Research System of RIKEN near Tokyo: "Squeeze it along its length. It can buckle either to the left or right." If shrunk to nanometer levels, the tape measure could adopt a superposition of twisting both left and right. In a paper submitted to *Physical Review Letters*, Nori and his colleagues Sergey Savel'ev and Xuedong Hu propose using either carbon nanotubes or bars machined from silicon as mechanical qubits.

A mechanical quantum computer would place molecular bars 10 to 30 nanometers long in rows spaced about 10 nanometers apart. Each bar would bear a charge, so that together their electric fields entangle their behavior, enabling the qubits to act in concert. The bars can get compressed either mechanically or electrically, and detection of each bar's state—the readout for the quantum computer—can be performed either optically or electrically.

It is too early to determine whether mechanical qubits can truly challenge other qubit approaches, such as superconducting circuits, which are also solid-state. "Superconductive devices have been worked on for about 40 years, and therefore much is known and many problems have been solved with these systems," points out physicist Andrew Cleland of the University of California at Santa Barbara. On the other hand, Cleland adds, the potential advantage of a mechanical system over an electronic system is that its qubits might intrinsically lose energy more slowly and thus remain in superposition longer, enabling them to perform more useful, complex calculations.

Nori and his co-workers plan to complete preliminary experiments on buckling nanotubes this year, with nanotubes in vacuum and at temperatures close to absolute zero to prevent interference from gas molecules or heat fluctuations. If they see the nanotubes in superposition as hoped, Nori guesses it will take one to three years to implement their mechanical qubits. He notes that the area seems to be advancing quickly—physicist Pritiraj Mohanty of Boston University and his colleagues in the January 28 *Physical Review Letters* described nanometer-size single-crystal beams of silicon that flicker between two different positions. Remarks Nori: "I can imagine Charles Babbage grinning right now."

Charles Q. Choi is a frequent contributor.

WHERE THE TWAIN SHALL MEET

Whether experiments on mechanical qubits, made out of bending nanotubes, will lead to a robust quantum computer is not yet certain. But being bigger than typical electronic and optical qubits, they should elucidate how the realm of atoms and molecules obeying quantum rules yields to the classical picture of everyday existence, where objects do not exist in two places at once. As Keith Schwab, senior physicist for the National Security Agency's laboratory at the University of Maryland, observes, "It could help to figure out where the border is between classical and quantum mechanics and what causes that boundary."

BY THE NUMBERS

The Lion's Share

MEASURING THE HUMAN IMPACT ON GLOBAL RESOURCES **BY RODGER DOYLE**

In recent years, environmental scientists have turned their attention to the total productive capacity of the earth and the share appropriated by *Homo sapiens*. Such measurements complement other indicators of global stress, such as atmospheric concentrations of greenhouse gases.

The latest and most detailed analysis comes from a group headed by biologist Marc L. Imhoff of the NASA Goddard Space Flight Center. Using satellite and climate data, the team calculated the worldwide amount of solar energy converted to plant organic matter through photosynthesis. This measure, called net primary production (NPP), is the chief source of food for living creatures. (It excludes other forms of energy, such as fossil fuels or seafood.) The researchers then figured out the human share of NPP based on consumption of plant foods, meat, milk, eggs, and wood used as fuel and in construction.

Imhoff and his colleagues took into account the efficiency of extraction and production methods, which varies considerably by country. In developed nations, for example, a metric ton of milled lumber requires 1.3 metric tons of tree biomass, compared with two metric tons in developing countries. The team derived the numbers for every 0.25 degree of latitude and longitude, equivalent to about 27.8 kilometers square at the equator.

The data show that average annual NPP from 1982 to 1998 was 119.6 billion metric tons of organic matter, of which humans appropriated 20 percent. (The uncertainties are large, resulting in a range of 14 to 26 percent.) The map indicates human appropriation of NPP as a percentage of that produced locally. The human share of NPP varies markedly by region. Large urban areas consume 300 times the amount of NPP generated locally, whereas in sparsely inhabited areas of the Amazon, human appropriation is close to zero.

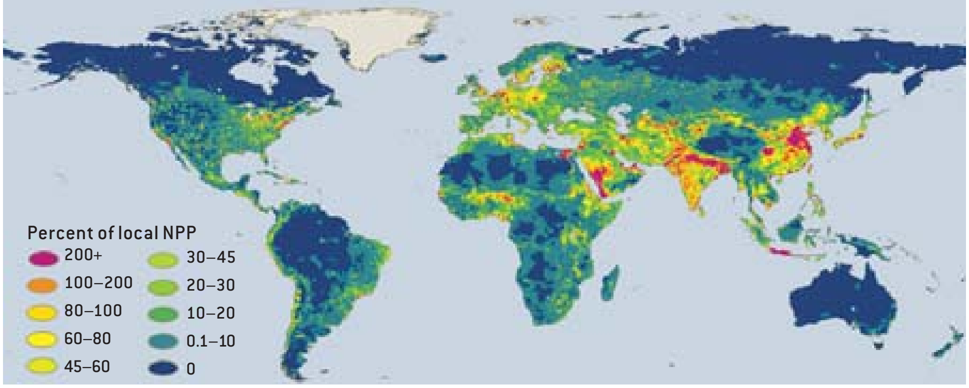
The increasing human share of plant and fiber could have worldwide implications, including for the composition of the atmosphere and hydrological cycles. It could also reduce biodiversity, which in turn could lead to declining NPP because of the disruption of such processes as crop pollination and pest control. The rate of species extinction already appears to be accelerating.

Preserving diversity is most difficult in the developing nations, where saving species is a distant second to filling the urgent needs of their people. Indeed, if these regions raised consumption to match that of industrial countries, human appropriation of NPP could rise to more than 35 percent.

Studies of NPP are at an early stage. If they become more precise in the future, such analyses will have great value in identifying the changing effect of consumption and technology and in evaluating policies aimed at reducing the human toll on ecosystems.

Rodger Doyle can be reached at rdoyle2@adelphia.net

HUMAN APPROPRIATION OF NET PRIMARY PRODUCTION (NPP)



RODGER DOYLE; SOURCE: MARC L. IMHOFF ET AL. (*map and data*)

FAST FACTS: PEOPLE'S SHARE

Percentage of human share of net primary production (NPP) in:

- South America: 6
- Africa: 12
- North America: 24
- East Asia: 63
- Western Europe: 72
- South Central Asia: 80

FURTHER READING

Spatial Sustainability, Trade and Indicators: An Evaluation of the "Ecological Footprint." Jeroen C.J.M. van den Bergh and Harmen Verbruggen in *Ecological Economics*, Vol. 29, No. 1, pages 61–72; April 1999.

Sharing the Garden. Christopher B. Field in *Science*, Vol. 294, pages 2490–2491; December 21, 2001.

Tracking the Ecological Overshoot of the Human Economy. Mathis Wackernagel et al. in *Proceedings of the National Academy of Sciences USA*, Vol. 99, No. 14, pages 9266–9271; July 9, 2002.

Global Patterns in Human Consumption of Net Primary Production. Marc L. Imhoff, Lahouari Bounoua, Taylor Ricketts, Colby Loucks, Robert Harriss and William T. Lawrence in *Nature*, Vol. 429, pages 870–873; June 24, 2004.

news SCAN

DATA POINTS: BANKRUPT HEALTH

Medical bills drive many personal bankruptcies in the U.S., according to a survey by a Harvard University team analyzing 1,771 cases in 2001. Its report appeared online February 2 in the journal *Health Affairs*.

Number of Americans who filed for bankruptcy in 2001: **1,458,000**

Percent of bankruptcies involving any medical cause: **54.5**

Percent caused specifically by*:

Illness or injury: **28.3**

Medical bills over $1,000: **27**

Loss of work income for medical reasons: **21.3**

New family member: **7.7**

Death in family: **7.6**

Alcohol or drug addiction: **2.5**

Mortgage to pay medical bills: **2**

Uncontrolled gambling: **1.2**

Percent of debtors who had insurance at outset of illness: **75.7**

Percent of medical debtors who, before going bankrupt:

Skipped doctor/dentist visit: **59.5**

Failed to fill a prescription: **46.7**

*Some respondents cited more than one specific cause.

SOURCE: www.healthaffairs.org

ATMOSPHERIC SCIENCE

Aurora Born of Radio

AURORA OVER ALASKA glows above an antenna array used to create artificial flashes.

The haunting glow of the aurora arises when energized electrons and ions excite gases in the atmosphere. Beams of radio waves can induce auroralike light, but emissions have been too faint to detect with the naked eye. Trying to make the best use of instrument time, Todd R. Pedersen of Hanscom Air Force Base in Massachusetts and Elizabeth A. Gerken of Cornell University trained their radio antennas on a natural aurora event in Alaska. They aimed for a lower than usual region of the atmosphere, about 100 kilometers high, where ions dissipate quickly compared with higher altitudes. For a few minutes last March small green speckles appeared amid the background glow—bright enough in principle to see directly, as the researchers describe in the February 3 *Nature*. If the effect proves reproducible, Pedersen says, the technique could serve as a tool for atmospheric studies. *—JR Minkel*

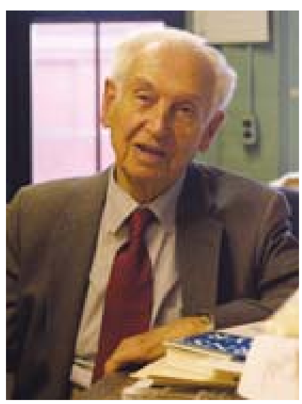
PSYCHOLOGY

Procrastinate Later

Ignoring the adage "Never put off until tomorrow what you can do today" is all too easy. Gal Zauberman of the University of North Carolina at Chapel Hill and John G. Lynch, Jr., of Duke University may have found out why: people idealize the future, expecting they will be less busy then. In surveys of 900 volunteers, they found that respondents could not gauge their future supply and demand of time as well as they could of money. If they lack knowledge of upcoming specific tasks, people act as if new demands will not inevitably arise that are as pressing as those that currently exist. When tomorrow changes into today, people discover they are too busy to do everything they promised. The findings were finished in time for the February *Journal of Experimental Psychology: General*. *—Charles Q. Choi*

OBITUARY

Ernst Mayr (1904–2005)

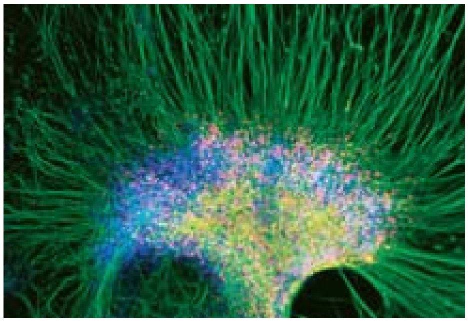


Calling him Darwin's leading defender, a profile of Ernst Mayr in the August 1994 issue of *Scientific American* noted that the then 90-year-old dean of evolutionists was writing a book called *This Is Biology*, to be followed by the simple volume *What Evolution Is*. Mayr saw the publication of those works as well as his 25th book, *What Makes Biology Unique?*, before his death on February 3 at his home in Bedford, Mass., after a brief illness. He was 100 years old. A member of the Harvard University faculty since 1953, Mayr was a naturalist, theorist, historian and philosopher whose books, some 700 journal articles and many lectures inspired and educated generations of researchers and the lay public. His July 2000 *Scientific American* essay, "Darwin's Influence on Modern Thought," was chosen for *The Best American Science Writing 2001;* it is available at www.sciam.com/ontheweb, as are a 100th-birthday interview, the 1994 profile and more. *—Steve Mirsky*

STEM CELLS

One Small Step

Researchers have turned human embryonic stem cells into spinal motor neurons, the long threadlike cells that carry signals



TRANSFORMED: Embryonic stem cells were coaxed to become neural cells, including motor neurons (*red*). Other features include neural fibers (*green*) and DNA in nuclei (*blue specks*).

from the spinal cord to the limbs and whose death paralyzes victims of spinal cord injury and disease. As an embryo develops in the womb, a changing mix of chemical factors prompts stem cells to morph into different kinds of neurons at specific times. A group led by Su-Chun Zhang at the University of Wisconsin–Madison hit on the combination of retinoic acid and a signaling molecule called sonic hedgehog to mimic the time when stem cells become spinal motor neurons, allowing them to grow the cells at will. The results, published in the February *Nature Biotechnology*, should enable scientists to look for drugs and develop treatments to counteract spinal cord disease. The late actor Christopher Reeve had lobbied for stem cell research for precisely this purpose. —*JR Minkel*

ENTOMOLOGY

Descent of the Ants

Rather than falling haphazardly, 25 species of arboreal ants in Panama, Costa Rica and Peru can glide back to their home trees—the first known instance of wingless insects guiding their fall. White nail polish on the ants' rear legs and high-speed video revealed that after being dropped from 30 meters up, ants can swivel quickly to glide backward to a tree. They can make 180-degree turns in midair that appear to involve abdominal undulations, airfoillike flattened hind legs, and flattened heads with flanges that can act as rudders. Eighty-five percent of canopy ant (*Cephalotes atratus*) workers returned to their home tree after falling, frequently crawling back to the branch from which they started within 10 minutes. Evidence suggests that the ants sometimes might purposely drop off trees to avoid predators. The report landed in the February 10 *Nature*.

—*Charles Q. Choi*

PUBLIC HEALTH

Crustaceans against Dengue Fever

Mosquitoes carrying dengue fever annually cause more than 50 million infections worldwide and at least 12,000 deaths. Striking at the bloodsuckers directly, Brian Kay of the Royal Brisbane Hospital in Australia and Vu Sinh Nam of the Ministry of Health in Hanoi sicced the crustacean *Mesocyclops* on the larvae of the mosquito *Aedes aegypti*. From 1998 to 2003 the tiny predator was spread in water tanks, and villages cleared away discarded containers likely to collect water. The result, presented in the February 12 *Lancet*, was the elimination of the mosquito in 32 Vietnamese communes and drastic reduction in five others. No dengue was reported in any of the 37 communes after 2002, meaning that some 380,000 people were protected. (The annual incidence rate before the program began ranged up to 146 cases per 100,000.) This solution might not prove effective everywhere, given the dislike many people have for visible organisms in the drinking water. —*Charles Q. Choi*

LARVAE HUNTER: Copepods have helped to prevent dengue fever.

BRIEF POINTS

- Neutron stars are born in the violence of a supernova, but one in a binary system may have formed when a white dwarf collapsed without exploding. A gentle birth would explain why the pair moves slowly.

  Physical Review Letters, February 11

- In mice, inflammation permits prions to enter organs not normally infected. It bolsters the theory that people with an existing illness may be more likely to get infected with mad cow prions when those brain-destroying particles are encountered.

  Science, February 18

- Foxes being bred in Siberia read human social cues as well as domestic dogs, even though the trait was not explicitly selected for. Evidently, social intelligence can rise simply when fear and aggression decrease.

  Current Biology, February 8

- An enzyme called kisspeptin, produced by the *kiss-1* gene, activates the puberty gene *GPR54*, a finding that may help those with hormonal development problems.

  Proceedings of the National Academy of Sciences USA, February 1

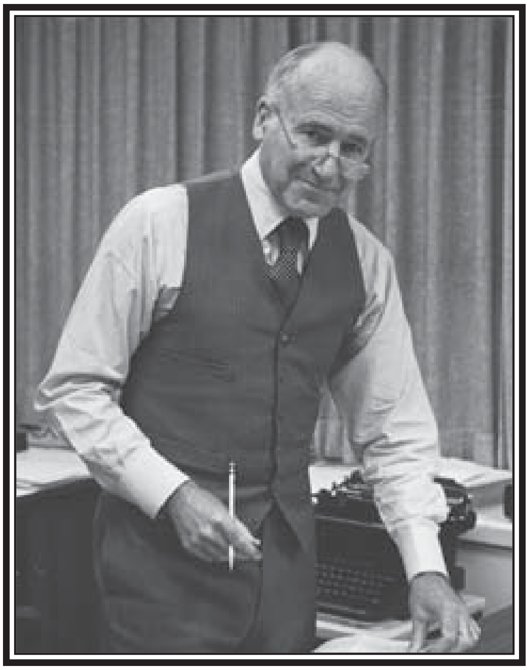
COURTESY OF SU-CHUN ZHANG (*top*); STEVE GSCHMEISSNER *Science Photo Library* (*bottom*)

IN MEMORIAM

Dennis Flanagan (1919–2005)

Dennis Flanagan, whose nearly four-decade tenure as editor of *Scientific American* transformed science journalism and educated untold millions about the wonders of science, passed away on January 14. The cause of his death was prostate cancer.

Flanagan's death falls only a few months after that of Gerard Piel, *Scientific American*'s former publisher and chairman. Flanagan and Piel, along with a small group of investors, purchased the magazine in 1947 and changed it from a rather quirky mix of science, inventions and mechanical hobbycraft into the world's premier voice of authoritative, intelligent science coverage. Both retired from active involvement with the magazine in the mid-1980s, when they sold it to its current owner, Verlagsgruppe Georg von Holtzbrinck.



DENNIS FLANAGAN in his office in the early 1980s.

Their partnership was an outgrowth of working together at *Life* magazine in the 1940s; Flanagan succeeded Piel as that magazine's science editor. *Scientific American* came to embody the enthusiasm and vision that the two men had for what a modern science magazine could be. Of course, not everyone appreciated Flanagan and Piel's *Scientific American* right off the bat. In the months after the magazine's May 1948 debut, it published a modest number of congratulatory letters, but also one that began, "Sirs: I have been trying to read the May issue of your magazine. Man, oh man. You have ruined the finest shop and hobby magazine in the world. Gone high-brow...."

Flanagan edited with a clear philosophy that "science is what scientists do," which lent the magazine its breadth of interests in all areas of research endeavor. Perhaps the best testament to the quality of his work (it surely gave Flanagan great pleasure) was a quote that has been attributed to the poet Robert Frost: "There are only two really great literary magazines in the United States. First is the *New Yorker*. The more brilliant of the two is *Scientific American*. It's come up with a great editorship. I'd rather read the advertisements in *Scientific American* than most of the literature written elsewhere."

In 1999 Flanagan was inducted into the American Society of Magazine Editors Hall of Fame.

Inspect the 37 years of issues that Dennis Flanagan edited, and you will look in vain for his byline. Flanagan never openly wrote a single article for his own publication. Yet his impeccable editorial judgment informed every line in those issues with clarity and intelligence.

As an author, Flanagan is best remembered for his book *Flanagan's Version,* a wide-ranging rumination on science that combines exposition, narrative and anecdote and that beautifully captures his genial astonishment at the natural world and at researchers' ingenuity in plumbing it.

A story about Flanagan immortalized in his book tells of the time he ran into film critic Pauline Kael at a luncheon and reproached her for boasting that she knew nothing of science. Kael reportedly replied, "Oh, you're a *Renaissance* hack!"

"It was a genial insult, and I had a good laugh," Flanagan wrote. "Actually I like the idea of being a Renaissance hack. If tombstones were still in style, I would want to have the two words chiseled right under my name. In an age of specialization people are proud to be able to do one thing well, but if that is all they know about, they are missing out on much else life has to offer."

Flanagan never missed a thing.

—The Editors

An expanded version of this memorial is available at www.sciam.com/ontheweb

RICHARD SASSO

The Feynman-Tufte Principle

A visual display of data should be simple enough to fit on the side of a van By MICHAEL SHERMER

I had long wanted to meet Edward R. Tufte—the man the *New York Times* called "the da Vinci of data" because of his concisely written and artfully illustrated books on the visual display of data—and invite him to speak at the Skeptics Society science lecture series that I host at the California Institute of Technology. Tufte is one of the world's leading experts on a core tool of skepticism: how to see through information obfuscation.

But how could we afford someone of his stature? "My honorarium," he told me, "is to see Feynman's van."

Richard Feynman, the late Caltech physicist, is famous for working on the atomic bomb, winning a Nobel Prize in Physics, cracking safes, playing drums and driving a 1975 Dodge Maxivan adorned with squiggly lines on the side panels. Most people who saw it gazed in puzzlement, but once in a while someone would ask the driver why he had Feynman diagrams all over his van, only to be told, "Because I'm Richard Feynman!"

EDWARD R. TUFTE, master of design analysis, poses next to a Feynman diagram on Feynman's van depicting the interaction of photons and electrons.

Feynman diagrams are simplified visual representations of the very complex world of quantum electrodynamics (QED), in which particles of light called photons are depicted by wavy lines, negatively charged electrons are depicted by straight or curved nonwavy lines, and line junctions show electrons emitting or absorbing a photon. In the diagram on the back door of the van, seen in the photograph above with Tufte, time flows from bottom to top. The pair of electrons (the straight lines) are moving toward each other. When the left-hand electron emits a photon (wavy-line junction), that negatively charged particle is deflected outward left; the right-hand electron reabsorbs the photon, causing it to deflect outward right.

Feynman diagrams are the embodiment of what Tufte teaches about analytical design: "Good displays of data help to reveal knowledge relevant to understanding mechanism, process and dynamics, cause and effect." We see the unthinkable and think the unseeable. "Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. Clear and precise seeing becomes as one with clear and precise thinking."

The master of clear and precise thinking meets the master of clear and precise seeing in what I call the Feynman-Tufte Principle: a visual display of data should be simple enough to fit on the side of a van.

As Tufte poignantly demonstrated in his analysis of the space shuttle *Challenger* disaster, despite the 13 charts prepared for NASA by Thiokol (the makers of the solid-rocket booster that blew up), they failed to communicate the link between cool temperature and O-ring damage on earlier flights. The loss of the *Columbia,* Tufte believes, was directly related to "a PowerPoint festival of bureaucratic hyperrationalism" in which a single slide contained six different levels of hierarchy (chapters and subheads), thereby obfuscating the conclusion that damage to the left wing might have been significant. In his 1970 classic work *The Feynman Lectures on Physics,* Feynman covered all of physics—from celestial mechanics to quantum electrodynamics—with only two levels of hierarchy.

Tufte codified the design process into six principles: "(1) documenting the sources and characteristics of the data, (2) insistently enforcing appropriate comparisons, (3) demonstrating mechanisms of cause and effect, (4) expressing those mechanisms quantitatively, (5) recognizing the inherently multivariate nature of analytic problems, (6) inspecting and evaluating alternative explanations." In brief, "information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory, skeptical."

Skeptical. How fitting for this column, opus 50 for me, because when I asked Tufte to summarize the goal of his work, he said, "Simple design, intense content." Because we all need a mark at which to aim (one meaning of "skeptic"), "simple design, intense content" is a sound objective for this series. SA

Michael Shermer is publisher of Skeptic *(www.skeptic.com) and author of* Science Friction.

MICHAEL SHERMER (*Tufte with Feynman van*); BRAD HINES

In the Business of Synthetic Life

Synthetic biology might someday lead to artificial organisms. To James J. Collins, it already offers pharmaceutical promise, like turning a person's cells into custom drug factories By SAM JAFFE

TRACY POWELL

JAMES J. COLLINS: MAKING LIFE

- **Practices synthetic biology, in which researchers tinker with genetic networks, rather than single genes of conventional genetic engineering.**
- **Found in previous work that a vibrating insole can improve the elderly's sense of balance, sparking interest from athletic shoe companies.**
- **On why engineering is easier than science: "All you have to do to succeed at engineering is to build something bigger, cheaper or faster. Science is creating new knowledge. That takes a lot more sweat and pain."**

At first glance, the bacterial colonies that dot a petri dish in the Boston University laboratory of James J. Collins do not seem all that special. Each *Escherichia coli* bacterium has been genetically altered to manufacture a specific protein once the population density of the colony around it reaches a predefined level.

A skeptic might yawn. After all, genetic engineering isn't new. But these cells haven't just had a foreign gene spliced into them. Collins inserted a whole genetic network—he put in many genes that interact together as well as with the natural genetic machinery of the cell. In this case, he dropped in a quorum-sensing network from a *Vibrio fischerii* bacterium. If conventional genetic engineering is like changing the blade on a screwdriver, then Collins's approach is akin to altering the contents of the entire toolbox at once.

The 39-year-old Collins is a member of an emerging field called synthetic biology. Practitioners create novel ingredients for the recipe of life, including nucleic acids, amino acids and peptides. Some of them even hope to manufacture an artificial organism [see "Synthetic Life," by W. Wayt Gibbs; SCIENTIFIC AMERICAN, May 2004]. It is still considered a seed-stage discipline, where brilliant young scientists wow one another with proof-of-concept experiments and publish papers filled with pages of mathematical formulas. Collins, on the other hand, is the first to generate commercial technologies that are in the advanced stages of development. More than any other, he is proving that synthetic biology is ready for the marketplace, much more quickly than others expected it could be.

The most promising of those technologies is an RNA ribo-regulator, which Collins first described in 2004. It consists of a sequence of DNA that, with the help of a genetically engineered virus, integrates into a host bacterium's genome. The DNA then creates a loop of messenger RNA that binds to a site on the ribosome (the cell's protein factory), thereby blocking the production of a specified protein. The regulator can do the opposite, too: it can unblock the ribosome on command in order to start making that protein. Essentially the ribo-regulator enables scientists to dictate protein production, with close to 100 percent accuracy and efficiency.

Others quickly improved on the ribo-regulator. Richard Mulligan of Harvard Medical School designed one that can be activated when a specific molecule is added to mouse cells. If these technologies prove successful inside humans, a person's cells could be turned into pharmaceutical plants. Pills would be popped only to turn the micro factories on or off. Such a future is still years away, but the progress thus far amazes Collins. "I never would have dreamed that within a year this technology would already be working in mammals," he says. A company founded by Collins, called Cellicon Biotechnologies, is now negotiating with several firms for use in drug discovery.

The ribo-regulator is not the only technology with such tremendous commercial promise coming from Cellicon. The company has encoded the principles behind synthetic biology into software to help screen drug candidates for their effect on the whole cell, rather than just on one protein target. "Drug companies are great at creating an assay that proves a compound hits a specific target," Collins states. "Thus far they haven't been very good at predicting what it will do to all the other genes and proteins in a cell."

By forward-engineering the cell, science could reveal its secrets more readily, Collins says.

Collins's success in technology development lies in the fact that he straddles the line between engineering and science so effortlessly. "I'm not sure if the conventional definitions are very helpful anymore," he says. "In the end, I'm far more interested in seeing the fruit of my work help a human being. If I do some good science along the way, that's great, too." Others agree. "Collins's scientific work is all the more impressive because he's done it while doing real engineering," remarks George Church, a biologist at Harvard Medical School.

Collins wanted to be an electrical engineer. But while attending the University of Oxford as a Rhodes scholar, Collins found himself studying nonlinear dynamics—popularly known as chaos theory—with Ian Stewart, the famed University of Warwick mathematician and former columnist for *Scientific American*'s Mathematical Recreations. "I haven't met anybody more emblematic of the concept of multidisciplinary research than Jim," Stewart says of his former protégé. "Some people just can't function without clear boundaries that define their discipline. Jim excels in such an environment."

After his work in the U.K., Collins became a faculty member of Boston University's department of bioengineering. He became intrigued with the relation between the human sense of balance and stochastic (or random) sensory inputs—better known as noise. "Normally you think of noise as hindering the clarity of a signal," Collins explains. "But in some cases, noise can enhance a signal." Collins hypothesized that senior citizens were losing their balance with age in part because they become less sensitive to stochastic stimuli, such as pressure on the soles of the feet. He designed a battery-powered shoe insole that produces just enough random vibrations to improve an average 75-year-old's sense of balance to that of a 25-year-old.

In the midst of working on the insole, he received an unusual request from Charles R. Cantor, his department chair. Cantor wanted Collins to use his expertise in nonlinear dynamics to make a presentation about genetic networks to a visiting grant-dispensing committee. Genes usually don't work alone. Instead they function within a system of interdependent genetic networks whose individual genes are constantly modifying the behavior of the other genes within the network. Collins, who at that point had no specialized education in molecular biology, spent the next four days preparing the presentation. The department didn't get the grant, but Collins realized from his crash course that biology was far closer to becoming an engineering discipline than most people realized. "Everybody was trying to reverse-engineer the cell, but that's the hardest way to understand it," Collins declares. "By forward-engineering it, science could reveal its secrets more readily."

Collins soon led a team that in 1999 created a genetic toggle switch. It consists of two foreign genes, each of which produces a protein that inhibits the other gene. Depending on the chemical added to the bacterial broth, the proteins of one gene would effectively be deactivated, disabling that gene. "The toggle switch is significant because no further modulation is necessary," Cantor says. Conventional genetic engineering needs continual insertion of a stimulant to keep the new gene running. The toggle switch stays on, or off, for as long as the organism remains alive.

Collins continues to optimize his toggle switch, which, like the ribo-regulator, has drawn interest from pharmaceutical companies. To some extent, the greatest promise of Collins's synthetic networks is that they help to verify the ever more complex software models that try to mimic the human cell. Yet Collins is adamant that such in silico modeling has its limits. "My holy grail isn't a virtual cell," he says, emphasizing a point that he believes the entire field of synthetic biology has to agree on to make further progress: "No matter how good we get at modeling, the model will never replace the actual experiment." SA

Sam Jaffe is a writer based in Philadelphia.

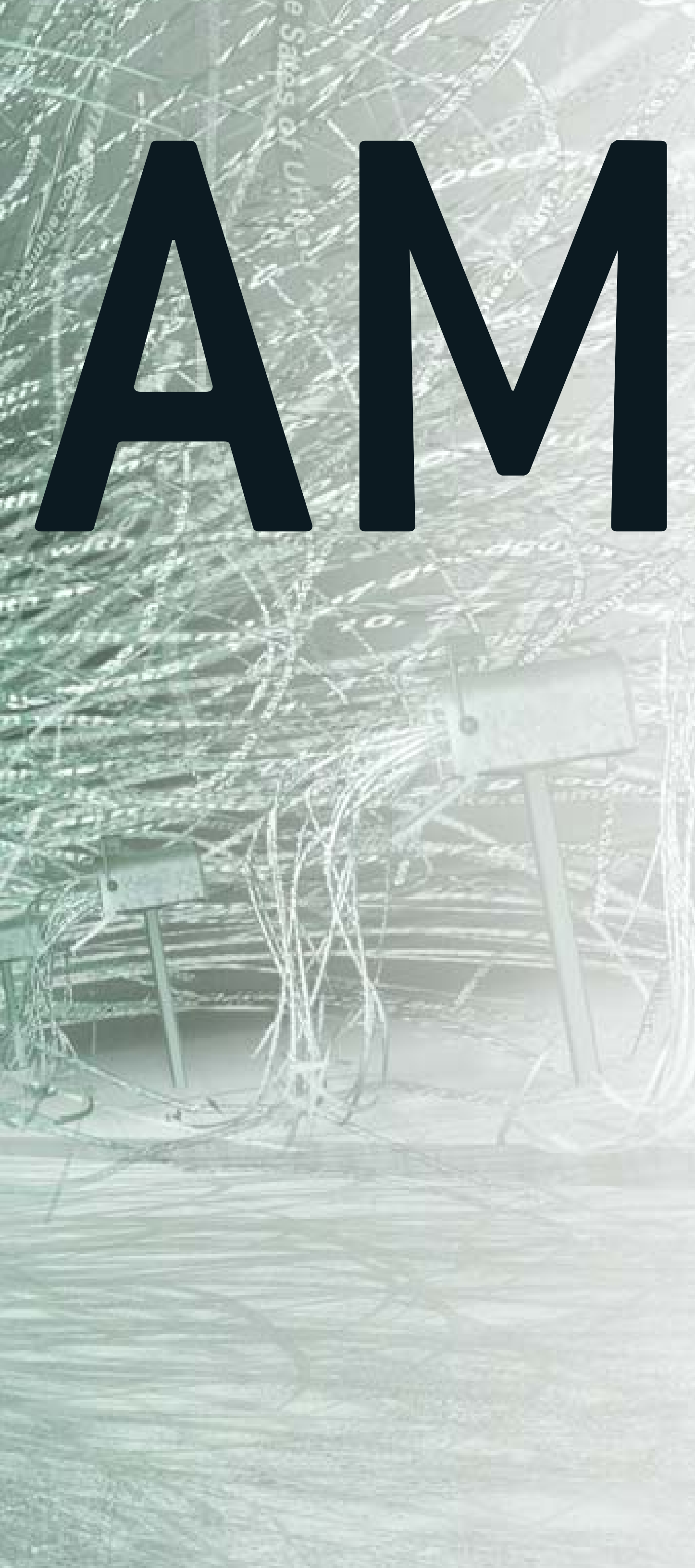
STOPPING SP

By Joshua Goodman, David Heckerman and Robert Rounthwaite

FOR ONE THIRD of all e-mail users, about four out of five incoming messages are spam. Computer scientists today are matched against spammers in an ongoing arms race to control what gets through to your in-box.

What can be done to stanch the flood of junk e-mail messages?

AM



In 1978 the first spam e-mail—a plug from a marketing representative at Digital Equipment Corporation for the new DEC-system-20 computer—was dispatched to about 400 people on the Arpanet. Today junk correspondence in the form of unwanted commercial solicitations constitutes more than two thirds of all e-mail transmitted over the Internet, accounting for billions of messages every day. For a third of all e-mail users, about 80 percent of the messages received are spam. Recently spam has become more threatening with the proliferation of so-called phishing attacks—fake e-mails that look like they are from people or institutions you trust but that are actually sent by crooks to steal your credit-card numbers or other personal information. Phishing attacks cost approximately $1.2 billion a year, according to a 2004 Gartner Research study.

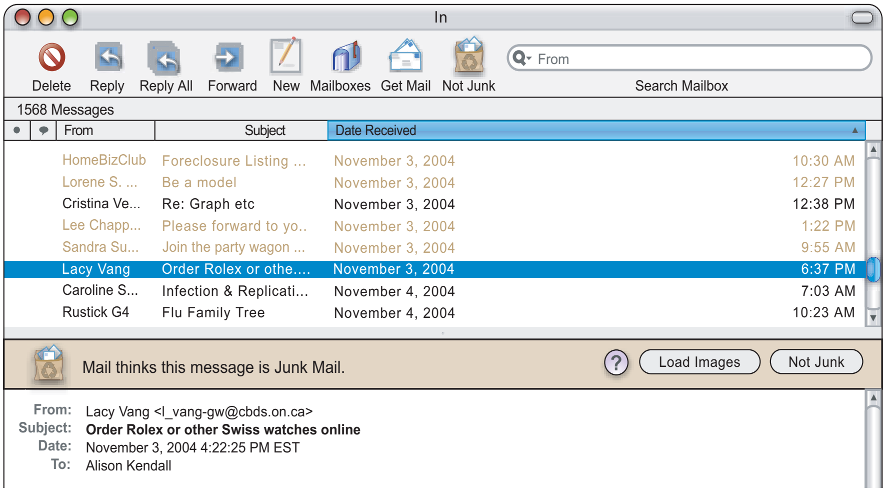
The phenomenon of spam afflicts more than just e-mail. Inside chat rooms lurk “robots” that pretend to be human and attempt to convince people to click on links that lead to pornographic Web sites. Instant messaging (IM) users suffer from so-called spIM—e-mail spam cognates. Blogs can be corrupted by “link spammers” who degrade Internet search engine operations by adding misleading links to sites that distort the utility ratings of Web sites and links.

The suffocating effect of spam sometimes seems likely to undermine, if not wreck, Internet communications as we have come to know them. The reality, however, is not so bleak. Several techniques for intercepting spam and discouraging spammers have been invented, and more are on the way. The methods we shall discuss focus on junk e-mail, but many of them could apply to other incarnations of spam as well. No one of these will be a magic cure, but combinations—if embraced by enough of us—could work wonders. It is not unrealistic to hope for a day when our e-mail boxes will once again be nearly spam-free.

Insidious E-mails

THE PROLIFERATION of fraudulent e-mail results directly from favorable market forces: spam is exceedingly cheap to distribute. It is not altogether free, though. We estimate that a message costs about one hundredth of a cent to send. At these cut-rate prices a spammer can earn only $11 per sale and still make a profit, even if the response rate is as low as one in 100,000. Hence, although very few e-mail users ever buy anything advertised in spam, all of us suffer because of those who do.

One of the most infuriating aspects of spam is that it changes continually to adapt to new attempts to stop it. Each time software engineers attack spam in some way, spammers find a way around their methods. This spam arms race has led to a continuous coevolution of the two, which has resulted in ever increasing sophistication on both sides.



GATEKEEPING SOFTWARE identifies spam e-mail by finding and assigning weights to features in previous examples that users have confirmed as wanted or unwanted.

Another fundamental problem stems from the fact that engineers and legislators find it extremely difficult to define spam. Most laws define it as unsolicited commercial e-mail from someone without a preexisting business relationship. This characterization is too broad, however. We recently received an e-mailed proposal, for example, to turn a short story we had published on the Internet into a motion picture. This communication met the requirements of the law: unsolicited, commercial, from an unknown sender, but almost no one would call it spam. An alternative definition might include the fact that spam is typically mass-mailed. But we recently solicited papers for a technical conference to discuss e-mail systems and anti-spam methods by sending requests to 50 people we had never met who had published on this topic. None of them complained. Perhaps the best characterization of spam is that it is poorly targeted and unwanted. Formulating a precise definition of spam is exceedingly difficult, but, like pornography, we certainly know it when we see it flooding our mailboxes.

Morphing Messages

WE HAVE WORKED ON the spam problem since 1997, when one of us (Heckerman) suggested that machine-learning methods might provide an effective line of attack. Since then, the three of us and our many colleagues in the software business have investigated and developed several approaches to stopping spam. They encompass combinations of technical and legal solutions as well as industrywide initiatives.

Some of the earliest schemes used to stop spam are so-called fingerprint-matching techniques. In these systems, engineers first find examples of spam and let computer programs "fingerprint" them. The fingerprint is a number derived from the content of the message, so that similar or identical messages get the same number. To give a simplified example, one could add the number of As in a message plus 10 times the number of Bs plus 100 times the number of Cs, and so forth. When a new message arrives, anti-spam programs compute its fingerprint and then compare it with those of known spam. If the fingerprints match, the program deletes or archives the message.

Regrettably, these straightforward methods were easily defeated by spammers, who simply started adding random characters to their messages [*see box on opposite page*]. Spam fighters responded with more sophisticated fingerprint techniques that try to exclude obvious sequences of random characters, but spammers overcame these efforts with more legitimate-looking random content, such as fake weather reports [*see box on opposite page*]. Ultimately, making fingerprint systems sufficiently robust to see through spammer randomizations turns out to be quite hard.

Smart Filters

RATHER THAN PURSUING fingerprint methods, our group followed an avenue that exploited machine-learning capabilities. These specialized computer programs can learn to distinguish spam e-mails from valid messages, and they are not so easily confused by additions of a few random letters or words.

At first, we tried the simplest and most common machine-learning method. The Naive Bayes algorithm starts with the probabilities of each word in the message. "Click," "here" and "unsubscribe," for instance, might each have a probability of 0.9 of showing up in spam and a probability of 0.2 of showing up in legitimate e-mail messages (1.0 being certainty). By multiplying the probabilities of all the words in a message and using a statistical principle known as Bayes' rule, we get an estimate of how likely it is that a message is spam.

The Naive Bayes strategy works remarkably well at determining what genuine e-mail looks like, and like all such learning methods, it resists simple obfuscations. Yet we were well aware of its shortcomings. Its assumption that words in e-mail are independent and unrelated

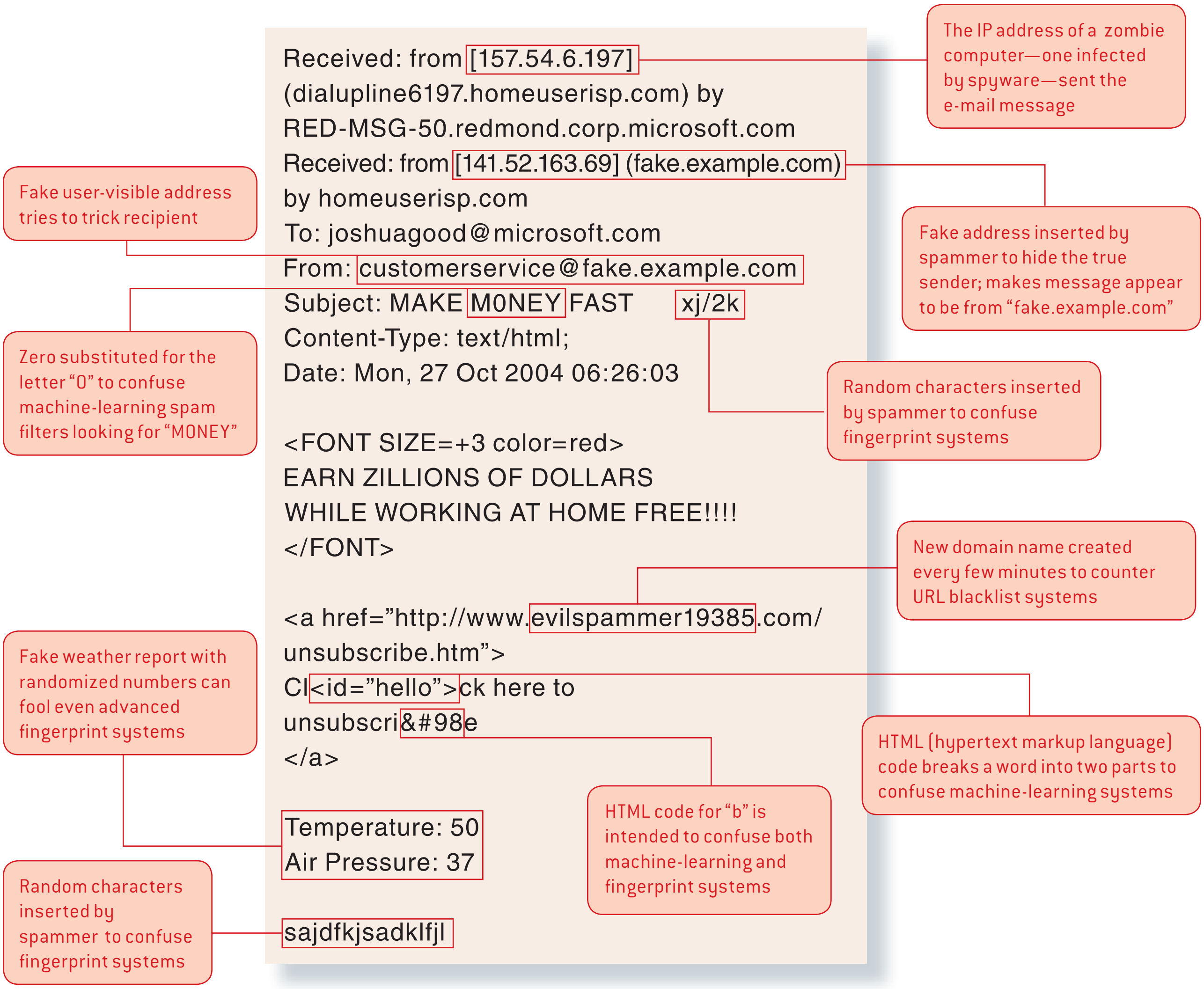
Overview/*Guarding Your In-Box*

- The growing tide of unwanted e-mail messages—spam—threatens the integrity of Internet communications. Software programmers are engaged in a continuing threat-versus-countermeasure battle against senders of junk messages, known as spammers.
- An array of existing and new anti-spam efforts, including smart software filters, systems that verify the legitimacy of e-mail senders and strong legal prohibitions, could stem the flow of spam if they were widely employed or enforced.

JEAN-FRANCOIS PODEVIN (*preceding pages*); ALISON KENDALL (*this page*)

SPAMMER PLOYS

Spammers use various methods to circumvent anti-spam filters. One of the most basic anti-spam measures is fingerprint matching, in which computers analyze known spam messages and then remove new messages that match them. Spammers soon learned to defeat simple fingerprint matching by adding sequences of random characters or random content such as fake weather reports, which change the fingerprint. Later, they began to obscure the spelling of words typically associated with spam—for example, by changing "MONEY" (with the letter "O") to "M0NEY" (spelled with a zero). Some filters also look for embedded links to known spam-related Web pages and servers. But spammers have learned to generate new addresses continually.



is in many cases false (for instance, "click" and "here" often appear together), which skews results.

Because of these difficulties, our research focuses on discriminative linear models, which optimize the models' later decisions when assigning weights to features. These features include words and properties of the messages, such as whether the message was sent to many recipients. These models can in some sense learn the relations between words—for instance, "knowing" not to place too much weight on words that tend to occur together, like "click," "here" and "unsubscribe." To explain further: let us say a Naive Bayes model saw these three words, which are often associated with spam. It might decide it has enough evidence to conclude that any message containing them is junk, leading it to sometimes delete valid e-mail. In contrast, a discriminatively trained model would know that the words tend to occur together and thus would assign lower, more reasonable, weights to them. Such a system could even learn that a word such as "here," which may occur more often in spam, should be given no weight at all because it does not really help tell good from bad. Discriminative methods can also discover that certain words cancel each other out. Although a word such as "wet" occurs more frequently in spam, when "wet" is found with "weather,"

ALISON KENDALL

chances are the message is legitimate.

An advantage of Naive Bayes systems is that they are easy to train. Determining the weights for discriminative methods is much harder: it requires programmers to try many sets of weight values for words and other features to find a combination that does the best overall job of distinguishing spam from nonspam. Fortunately, researchers have made significant progress here. Algorithms such as the Sequential Minimal Optimization algorithm, invented by John C. Platt of Microsoft, and the Sequential Conditional Generalized Iterative Scaling (SCGIS) algorithm, created by one of us (Goodman), are tens or hundreds of times faster than older techniques. When dealing with large amounts of spam training data, more than a million messages and hundreds of thousands of weights, quicker algorithms are critical.

Hiding Spam

WE HAD ALWAYS KNOWN that our machine-learning systems, which focus on the words in a message, would be vulnerable to spammers who obscure the wording of their output. Clever spammers, for example, learned to use words such as "M0NEY" (with a zero instead of the letter "O") or to use HTML (hypertext markup language) tricks, such as splitting a word into multiple parts (say, "cl" and "ick" instead of "click"). Because the telltale terms ("money," "click") are no longer in the message, the filter can be confused [*see box on preceding page*]. The good news is that machine-learning systems can often learn about these tricks and adapt.

Unfortunately, we had assumed erroneously that few people would respond to a message that was obviously attempting to defeat a spam filter—for we thought, who would buy a product like that? Sadly, we were wrong; purchasers of illicit or illegal products do not expect the sellers to employ respectable advertising techniques. So we have had to alter our learning systems by employing what researchers call n-gram models. These techniques use subsequences of words to detect the key words often associated with spam. If an e-mail message contains the phrase "n@ked l@dies", for instance, the n-grams extracted from this phrase would include "<space>n@k," "n@ke," "@ked," and so on. Because these word fragments appear in confirmed spam messages, their presence provides valuable clues.

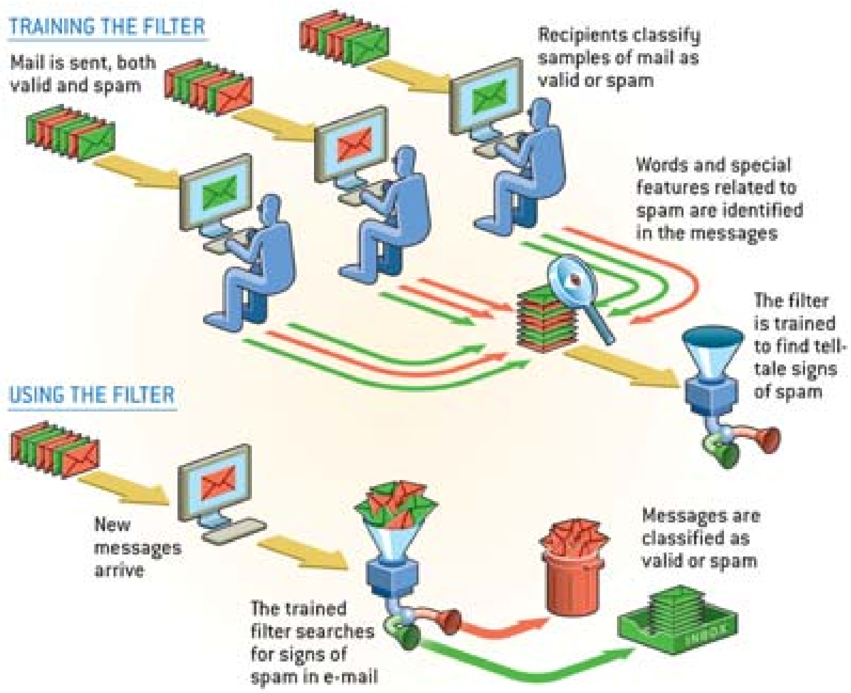
N-gram techniques have also helped us improve the utility of our filters when they are applied to foreign languages. Japanese and Chinese, for example, do not use spaces to separate words, so explicitly finding word breaks is very difficult. For these languages, n-gram-enabled systems simply screen every possible word and word fragment.

Image-Based Spam

SPAMMERS SOMETIMES HIDE their message in an image, where machine-learning systems cannot analyze the content (although they can still exploit other clues, such as the links in the message, sender reputation information, and so forth). One promising area of future

E-MAIL USERS TRAIN SPAM FILTERS

The first spam filters that Microsoft shipped processed data collected from just 20 e-mail users. As spammers grew more sophisticated, a better data source was needed. Today 100,000 Hotmail volunteers help to gather junk messages. These users classify a random selection of their own messages as legitimate or spam, which lets the system learn new targets to filter on a regular basis. When spammers find a way to circumvent the filter, it takes only a short while until the system determines how to recognize the new spam messages and screen them out.



SAMUEL VELASCO

THE AUTHORS

JOSHUA GOODMAN, DAVID HECKERMAN and *ROBERT ROUNTHWAITE* have worked together on ways to stop spam for many years. Heckerman and Rounthwaite, with others, created the first machine-learning spam filter in 1997. Heckerman manages the Machine Learning and Applied Statistics (MLAS) group at Microsoft Research. Goodman and Rounthwaite helped to organize the Microsoft product team that delivers the anti-spam technologies deployed in Exchange, Outlook, MSN and Hotmail. Rounthwaite is currently the group's chief architect. Goodman is a member of the MLAS team and does research on spam and e-mail-related topics.

research is the use of optical character-recognition (OCR) techniques for spam filtering. The same OCR techniques that are used for scanning a document could find all the text in the images and feed it to a machine-learning filter.

One of the more offensive aspects of spam is the appearance of pornographic images in one's mailbox. Fortunately, computer-vision researchers have made great progress in the automatic detection of pornographic images. Work in this field is surprisingly broad, for it has applications in preventing children's access to Web sites containing sexual material and in preventing pornographers from abusing free Web hosting systems. Such image recognition is, however, still time-consuming, and the reliability of identification needs to improve. Benign images, especially those showing large amounts of skin, can trigger false positives [*see box on next page*].

Our team is also investigating the analysis of universal resource locator (URL) information—the code that links to Web pages—to distinguish spam. Ninety-five percent of spam messages contain a URL. Most spammers' first goal is to get users to visit their Web site (although a small fraction prefer contact through telephone numbers), so URL information is an especially good target for filters.

Filters can exploit URL information in many ways. Some anti-spam software providers have already started blocking spam that contains links to known spam-related Web pages. Links to previously unknown domains can be considered suspicious: spammers generate new domains very quickly, whereas most legitimate domains are long-lived. On the other hand, URL information can also be an indicator of legitimate e-mail: a message that contains only pointers to known non-spam-related pages, or no URLs at all, is much less likely to be spam.

Prove It

ALTHOUGH FILTERING techniques work quite well, we recognize that spammers will always try to attack them. Rather than trying to win this endless competition, we believe the most effective approach in the long run would be to change the rules of the game. Hence, we are exploring proof systems—those whose goal is to require more from a spammer than he or she can afford.

That very first spam message was sent by manually typing in all 400 e-mail addresses. Today nearly all spam is sent automatically. If a sender can prove he or she is a human being, therefore, the sender is probably not a spammer. One of the earliest proof systems, suggested by Moni Naor of the Weizmann Insti-

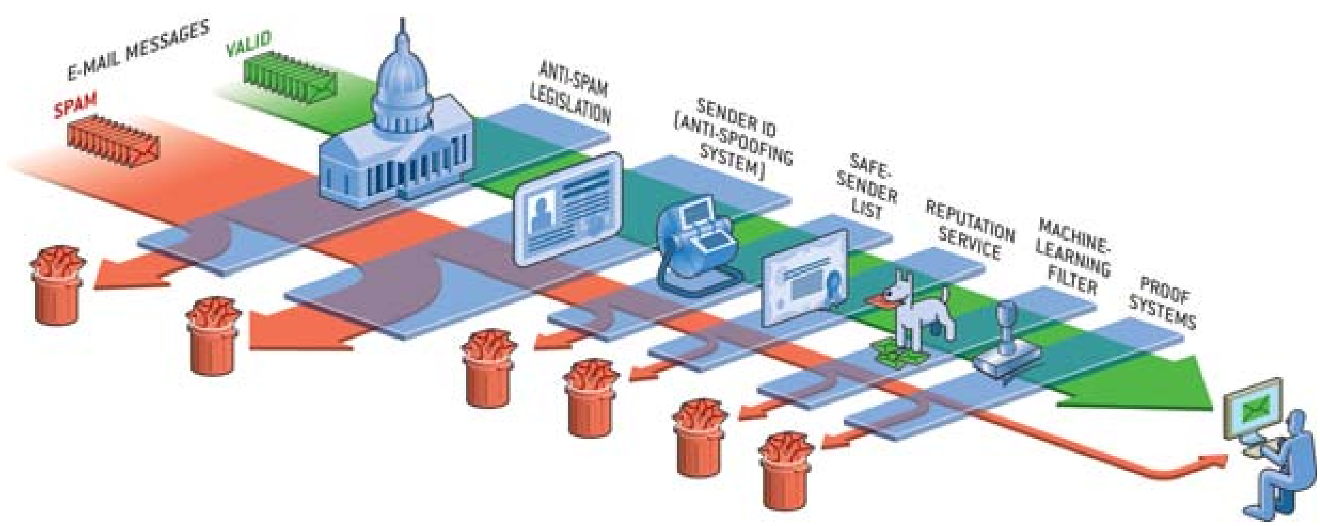
MULTIPLE ANTI-SPAM EFFORTS

Halting spam may require a multipronged defense that combines laws that discourage spam, technologies that find fake sender addresses, smart e-mail filtering and proof systems that confirm that a sender is a person or that make it too costly for spammers to operate.

The first line of defense is governmental legislation. The CAN-SPAM act prohibits certain particularly nasty practices, although so far spammers seem mostly undeterred.

Because about half of spam messages use fake sender addresses, the new Sender ID Framework standard adds supplementary information to the domain name server (DNS), listing the Internet protocol (IP) addresses of computers allowed to send mail from that domain.

The e-mail system in a PC checks a "safe list" of trusted senders preselected by the recipient. Senders on the list bypass further filtering. Those not on the personal list would then be checked by a reputation service for senders who have agreed to abide by strict standards that prevent spam from being sent. If a sender was not on either list, the message would go to a machine-learning-based anti-spam filter. Senders that seem even slightly suspect would be required to provide a form of proof—such as solving a simple puzzle that confirms the sender is a person, answering a more difficult puzzle that requires costly computer processing time, or submitting a small, refundable payment.



SAMUEL VELASCO

tute of Science in Israel, made use of this notion. Naor proposed using what became known variously as HIPs (human interactive proofs), CAPTCHAs—an acronym for "completely automated public Turing test to tell computers and humans apart"—or reverse Turing tests [see "Baffling the Bots," by Lee Bruno; SCIENTIFIC AMERICAN, November 2003]. A HIP is a problem or puzzle designed to be easy for most humans but as difficult as possible for computers. People, for instance, are far superior to machines at recognizing sets of random alphabet letters that are partially obscured or distorted in an image.

A HIP forms part of a challenge-response system, which verifies that the sender is human. Before delivering a message, the system first checks a "safe list" of senders that the recipient considers trustworthy. If the sender is on the list, the message is delivered to the recipient's mailbox. If not, a challenge message goes to the original sender asking him or her to solve a HIP. After the sender solves the HIP, the response travels back to the recipient, whose e-mail software then transfers the message to the recipient's in-box.

This kind of interactive system can be annoying to users, however. Few people want to solve HIPs to send e-mail messages, and some even refuse to do so. An automated alternative proof mechanism, suggested by Naor and his colleague Cynthia Dwork, uses computational puzzles. To deliver a message successfully, the sender's e-mail system must first work out a computational puzzle posed by the recipient's system. The idea is to prove that the sender has expended more computer time on that individual message than a mass-marketing spammer could afford. Computational puzzles are like jigsaw puzzles—difficult to solve but easy to verify. On average, they could require many seconds or even minutes to find a solution but only milliseconds to validate. Solving these problems promptly would require spammers to buy many computers, making their costs prohibitive.

Yet another kind of proof system uses real money. Senders include with their message a kind of electronic check for a small amount, say a penny. Including the check allows their message through spam filters. If the message is good, the recipient ignores the check, but if the message proves to be spam, a standardized complaint mechanism allows the recipient to cash it (or donate it to charity). Rate-limiting software meanwhile monitors senders' message volumes, ensuring they do not send more mail than their balance allows. For legitimate senders, this system is free, but for spammers, the cost per message might be one cent, 100 times our estimate of the current price—more than spammers can afford. For individuals, a small virtual deposit is also made by their Internet service provider or when they purchase e-mail software, so that for most users there is no cost at all.

Though straightforward in concept, monetary systems of this kind will be difficult to put into practice. Electronic systems require some overhead, so these transactions will not be free. Many questions about a micropayment banking infrastructure remain unanswered: Where will the money to pay for it come from? How will its operations be sustained, and who will profit? Who will get the payments, and how will the system prevent fraud? Although none of these problems are insoluble, setting up such a scheme will be tough.

PORNOGRAPHIC Image-Recognition Systems

Automated anti-pornographic systems that filter out sexually explicit images sometimes mistake benign images, such as those above, for more objectionable ones.

JOSHUA GOODMAN (*left*); DAN GURAVICH *Corbis* (*right*)

All-Inclusive Attack

OUR FAVORITE STRATEGY to halt spam combines e-mail filtering technology with a choice of proof tests: HIPs, computational puzzles and micropayments. In this approach, if the sender of a message is not on the recipient's safe list, the message is shunted to a machine-learning-based anti-spam filter that is designed to be especially aggressive; if the message is even a bit suspicious, the recipient is challenged. Most messages from one person to another, however, will not be contested, which reduces the number of proofs dramatically. The original sender is then given a choice: solve a HIP or a computational puzzle or make a refundable micropayment. If the sender's computer has newer software, it will work out the puzzle automatically, without the sender even being aware of the challenge. Otherwise, the sender will solve a HIP or make a micropayment.

Of course, individual companies or institutions, no matter how large, can make only so much progress against spam. A comprehensive solution requires cooperation of the entire computer and software industry, as well as national governments.

Approximately two thirds of all e-mail today uses "spoofed," or fake, sender addresses. The e-mail protocols in use today are based on trust: senders simply state who they are and the recipients believe them. This approach worked quite well in the early days of the Internet, before spam proliferated and before e-mail was used for business transactions.

Changing Internet standards is notoriously difficult, and it has been especially hard for e-mail protocols. A new industry standard, the Sender ID Framework, is finally addressing the spoofing problem, however. It works by adding supplementary information to the domain name server (DNS) to list the Internet protocol (IP) addresses from which mail sent from a specific domain (part of the network) can come. IP addresses are numeric addresses, like street addresses for individual computers, such as "1.2.3.4." The new DNS list of entries for a given domain—say, "example.com"—determines which IP addresses are allowed to send mail from that domain. If a spammer pretends to be example.com, his or her IP address will not match any IP address in example.com's Sender ID entries, and an e-mail program will know the spammer's mail is fake.

Although knowing the identity of the sender is a critical step in preventing fraud (such as phishing e-mails), it will not solve the spam problem. Nothing stops spammers from making up new identities every day or even every few minutes. That is why reputation services—by which senders can certify themselves as legitimate—will be so important.

In one case, IronPort's Bonded Sender program, senders deposit money as a bond surety. If complaint rates from the sender exceed a certain threshold, bond money is forfeited to a specified charity. Spam filters can check the Bonded Sender list and allow mail from a certified sender past the spam filter, even if it seems suspicious. Such programs can work even for those who send few messages. An Internet service provider (ISP) such as MSN or AOL, for example, might join a reputation service to gain access to its certification program; the ISP would then monitor each of their users' e-mail volume and complaint rates, thus ensuring that none of the provider's users are spammers.

If most legitimate senders adopted such a system (and there is little reason why they would not), spam filters could be made to be much more aggressive in dealing with the remaining mail, thus stopping the vast majority of junk messages. Reputation systems could be combined with challenge-response systems, so that those who cannot join have an alternative method for sending mail.

A complementary approach to stopping spam is governmental legislation. The CAN-SPAM act went into effect in the U.S. in January 2004. The act itself does not outlaw spamming; it only prohibits certain particularly egregious techniques, such as using fake "From:" information. Unfortunately, CAN-SPAM has had little measurable effect so far. The proportion of spam with a fraudulent "From:" address has actually increased from 41 to 67 percent since the act went into effect. European nations, in contrast, have passed much stricter opt-in laws, which prevent people from sending commercial e-mails without explicit permission from the recipient. According to anecdotal evidence, these laws have been somewhat effective, at least in stopping spamming by large legitimate companies.

Clearly, no law of a single country can hope to end spam. Only about half of all junk e-mail comes from the U.S.; the rest originates overseas. Only about one in three products sold via spam (such as insurance or mortgage refinancing) requires a domestic U.S. presence. Others, including pornography, "herbal enhancers" and confidence scams, are already abroad, can easily move offshore or are already illegal.

Spam-Free Future

INDUSTRY, the open-source community and the academic community all continue to study how to eliminate spam. We recently helped to establish the first formal conference on the topic—the Conference on Email and Anti-Spam, which attracted researchers from all over the world. Engineers at IBM showed how to use techniques from bioinformatics, originally designed for finding patterns in genes, to discern patterns in spam. AOL investigators demonstrated that multiple fingerprint systems with different vocabularies could better defend against spammer obfuscations. A team from the University of California at Davis described how the addition of a few common words could produce an effective attack against machine-learning spam filters and how, with training, the filters could be made more resistant to this attack.

We have little doubt that the combination of the current and next-generation techniques will eventually stop most spam. There will always be a few spammers, of course, who are willing to pay the price to get through to our mailboxes, but the flood will turn into a trickle. SA

MORE TO EXPLORE

A Bayesian Approach to Filtering Junk E-Mail. M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. AAAI Technical Report WS-98-05, Madison, Wis., 1998. **http://citeseer.ist.psu.edu/sahami98bayesian.html**

Pattern Classification. Second edition. Richard O. Duda, Peter E. Hart and David G. Stork. John Wiley & Sons, 2000.

Learning to Filter Unsolicited Commercial E-Mail. Ion Androutsopoulos, Georgios Paliouras and Eirinaios Michelakis. Technical Report 2004/2, NCSR Demokritos. **http://iit.demokritos.gr/~paliourg/papers/TR2004_updated.pdf**

Spam Kings: The Real Story behind the High-Rolling Hucksters Pushing Porn, Pills, and %*@)# Enlargements. Brian McWilliams. O'Reilly, 2004.

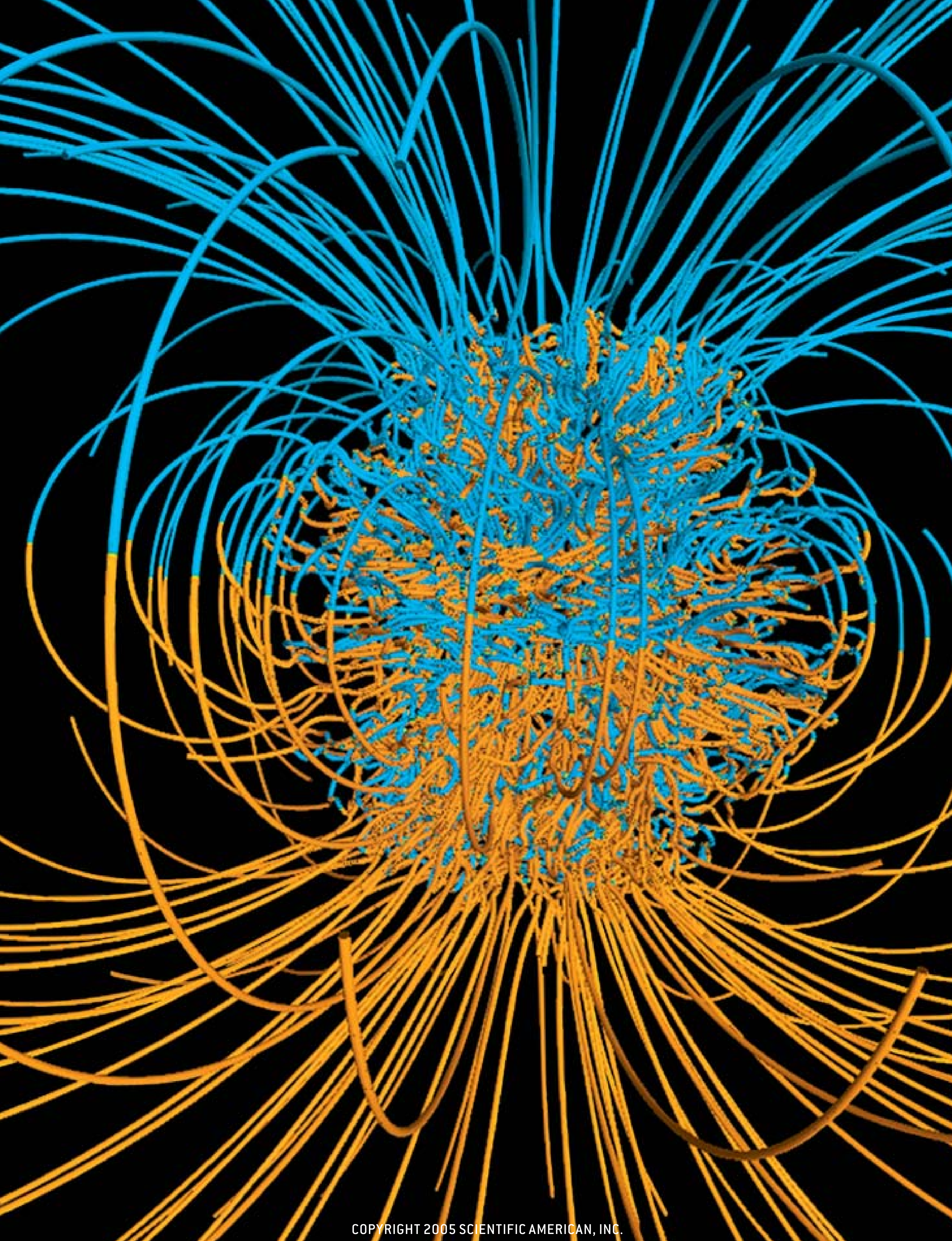
Conference on Email and Anti-Spam: **www.ceas.cc**

A Plan for Spam. Paul Graham. **www.paulgraham.com/spam.html**

Spam: Technologies and Policies. Joshua Goodman. **www.research.microsoft.com/~Joshuago/spamtech.pdf**

Tips for consumers on how to avoid spam: **www.microsoft.com/athome/security/spam/fightspam.mspx**

U.S. Federal Trade Commission's Web site on spam-related issues: **www.ftc.gov/spam/**



GARY A. GLATZMAIER

Probing the Geodynamo

Scientists have long wondered why the polarity of the earth's magnetic field occasionally reverses. Recent studies of our planet's churning interior are offering intriguing clues about how the next reversal may begin

BY GARY A. GLATZMAIER AND PETER OLSON

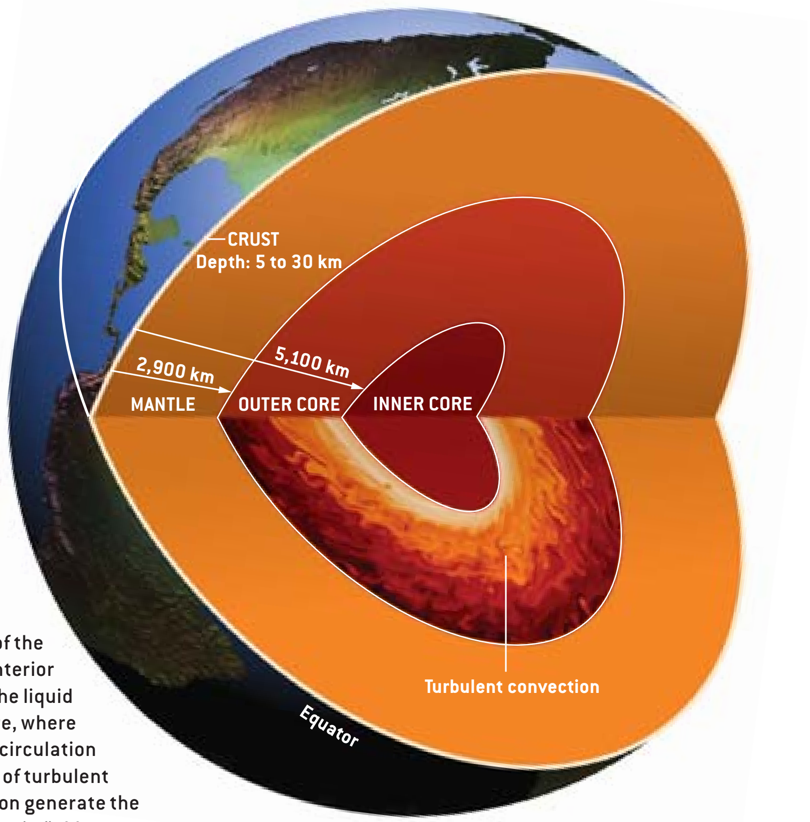
Most of us take it for granted that compasses point north. Sailors have relied on the earth's magnetic field to navigate for thousands of years. Birds and other magnetically sensitive animals have done so for considerably longer. Strangely enough, however, the planet's magnetic poles have not always been oriented as they are today.

Minerals that record past orientations of the earth's magnetic field reveal that it has flipped from north to south and back again hundreds of times during the planet's 4.5-billion-year history. But a switch has not occurred for 780,000 years—considerably longer than the average time between reversals, about 250,000 years. What is more, the primary geomagnetic field has lessened by nearly 10 percent since it was first measured in the 1830s. That is about 20 times faster than the field would decline naturally were it to lose its power source. Could another reversal be on its way?

Geophysicists have long known that the source of the fluctuating magnetic field lies deep in the center of the earth. Our home planet, like several other bodies in the solar system, generates its own magnetic field through an internal dynamo. In principle, the earth's dynamo operates like the familiar electric generator, which creates electric and magnetic fields from the kinetic energy of its moving parts. In a generator, the moving parts are spinning coils of wire; in a planet or star, the motion occurs within an electrically conducting fluid. A vast sea of molten iron more than six times the volume of the moon circulates at the earth's core, constituting the so-called geodynamo.

Until recently, scientists relied primarily on simple theories to explain the geodynamo and its magnetic mysteries. In the past 10 years, however, researchers have developed new ways to explore the detailed workings of the geodynamo. Satellites are providing clear snapshots of the geomagnetic field at the earth's surface, while new strategies for simulating earthlike dynamos on supercomputers and creating physical models in the laboratory are elucidating those orbital observations. These efforts are providing an intriguing explanation for how polarity reversals occurred in the past and clues to how the next such event may begin.

MAGNETIC LINES OF FORCE from a computer simulation of the geodynamo illustrate how the earth's magnetic field is much simpler outside the planet than inside the core (*tangled tubes at center*). At the earth's surface, the main part of the field exits (*long yellow tubes*) near the South Pole and enters (*long blue tubes*) near the North Pole.



DISTINCT LAYERS of the earth's interior include the liquid outer core, where complex circulation patterns of turbulent convection generate the geomagnetic field.

JEN CHRISTIANSEN (*cutaway*); GARY A. GLATZMAIER (*outer core overlay*)

Driving the Geodynamo

BEFORE WE EXPLORE how the magnetic field reverses, it helps to consider what drives the geodynamo. By the 1940s physicists had recognized that three basic conditions are necessary for generating any planet's magnetic field, and since then other findings have built on that knowledge. A large volume of electrically conducting fluid, the iron-rich liquid outer core of the earth, is the first of these conditions. This critical layer surrounds a solid inner core of nearly pure iron that underlies 2,900 kilometers of solid rock that form the massive mantle and the ultrathin crust of continents and ocean floors. The overlying burden of the crust and mantle creates average pressures in the core two million times that at the planet's surface. Core temperatures are similarly extreme—about 5,000 degrees Celsius, similar to the temperature at the surface of the sun.

These extreme environmental conditions set the stage for the second requirement of planetary dynamos: a supply of energy to move the fluid. The energy driving the geodynamo is part thermal and part chemical—both create buoyancy deep within the core. Like a pot of soup simmering on a burner, the core is hotter at the bottom than at the top. (The core's high temperatures are the result of heat that was trapped at the center of the earth during its formation.) That means the hotter, less dense iron in the lower core tends to rise upward like blobs of hot soup. When the fluid reaches the top of the core, it loses some of its heat in the overlying mantle. The liquid iron then cools, becoming denser than the surrounding medium, and sinks. This process of transferring heat from bottom to top through rising and sinking fluid is called thermal convection.

In the 1960s Stanislav Braginsky, now at the University of California at Los Angeles, suggested that heat escaping from the upper core also causes the solid inner core to grow larger, producing two extra sources of buoyancy to drive convection. As liquid iron solidifies into crystals onto the outside of the solid inner core, latent heat is released as a by-product. This heat contributes to thermal buoyancy. In addition, less dense chemical compounds, such as iron sulfide and iron oxide, are excluded from the inner core crystals and rise through the outer core, also enhancing convection.

For a self-sustaining magnetic field to materialize from a planet, a third factor is necessary: rotation. The earth's rotation, through the Coriolis effect, deflects rising fluids inside the earth's core the same way it twists ocean currents and tropical storms into the familiar spirals we see in weather satellite images. In the core, Coriolis forces deflect the upwelling fluid along corkscrewlike, or helical, paths, as though it were following the spiraling wire of a loose spring.

That the earth has an iron-rich liquid core, sufficient energy to drive convection and a Coriolis force to twist the convecting fluid are primary reasons why the geodynamo has sustained itself for billions of years. But scientists need additional evidence to answer the puzzling questions about the magnetic field that emerges—and why it would change polarity over time.

Overview/*Turbulence Matters*

- The geologic record reveals that the earth's primary magnetic field switches polarity every so often, and researchers have long wondered why.
- Recent computer models of fluid motion in the earth's molten core have simulated an earthlike magnetic field and associated polarity reversals. But because the fluid motion in these models is considerably simpler than the turbulent patterns thought to exist inside the earth, it is unclear how true to life these findings really are.
- Three-dimensional models capable of simulating turbulence, which are now under development, will one day resolve some of that uncertainty. In the meantime, satellite maps of the magnetic field and laboratory convection experiments are providing additional insight.

Magnetic Field Maps

A MAJOR DISCOVERY unfolded over the past five years as it became possible for scientists to compare accurate maps of the geomagnetic field taken 20 years apart. A satellite called Magsat measured the geomagnetic field above the earth's surface in 1980; a second satellite—Oersted—has been doing the same since 1999 [*see illustration on page 55*]. Investigators have mathematically projected these satellite measurements down to the top of the core using the assumption that the electric currents of the earth's mantle are negligible. The core-mantle boundary is the closest researchers can get to the much more intense and complicated magnetic field that exists within the core, where magnetic fluctuations actually originate; strong electric currents in the core prevent direct measurements of the magnetic field there. Despite the inherent limitations, several noteworthy observations came out of these efforts, including hints about the possible onset of a new polarity reversal.

One important finding was that most of the geomagnetic field originates at only four broad regions on the core-mantle boundary. Although the geodynamo produces a very intense magnetic field, only about 1 percent of the field's magnetic energy extends outside the core. When measured at the surface, the dominant structure of this field is called the dipole, which most of the time is roughly aligned with the earth's axis of rotation. Like a simple bar magnet, this field's primary magnetic flux is directed out from the core in the Southern Hemisphere and down toward the core in the Northern Hemisphere. (Compass needles point to the earth's north geographic pole because the dipole's south magnetic pole lies near it.) But the satellite missions revealed that the flux is not distributed evenly across the globe. Instead most of the dipole field's overall intensity originates beneath North America, Siberia and the coast of Antarctica.

Ulrich R. Christensen of the Max Planck Institute for Solar System Research in Katlenburg-Lindau, Germany, suspects that these large patches come and go over thousands of years and stem from the ever evolving pattern of convection within the core. Might a similar phenomenon be the cause of dipole reversals? Evidence from the geologic record shows that past reversals occurred over relatively short periods, approximately 4,000 to 10,000 years. It would take the dipole nearly 100,000 years to disappear on its own if the geodynamo were to shut down. Such a quick transition implies that some kind of instability destroys the original polarity while generating the new polarity.

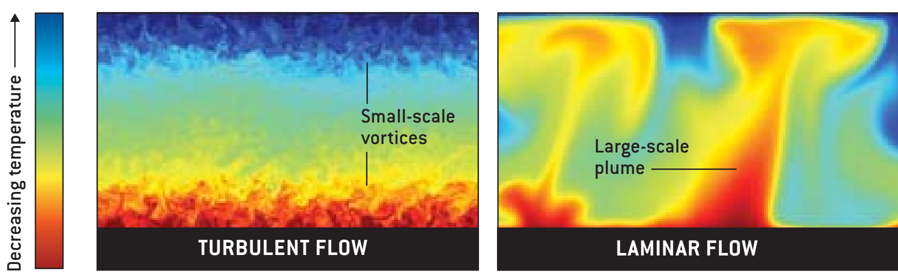
In the case of individual reversals, this mysterious instability is probably some kind of chaotic change in the structure of the flow that only occasionally succeeds in reversing the global dipole. But the frequency of reversals, which has been increasing steadily for the past 120 million years [*see illustration on page 57*], may have an external control. One possible candidate is a change in temperature at the bottom of the mantle, which could force the core to change its upwelling patterns.

Symptoms of a possible reversal-inducing change came to light when another group analyzed the Magsat and Oersted satellite maps. Gauthier Hulot and his colleagues at the Geophysical Institute in Paris noticed that sustained variations of the geomagnetic field come from places on the core-mantle boundary where the direction of the flux is *opposite* of what is normal for that hemisphere. The largest of these so-called reversed flux patches stretches from under the southern tip of Africa westward to the southern tip of South America. In this patch, the magnetic flux is inward, toward the core, whereas most of the flux in the Southern Hemisphere is outward.

Patch Production

ONE OF THE MOST significant conclusions that investigators drew by comparing the recent Oersted magnetic measurements with those from 1980 was that new reversed flux patches continue to form on the core-mantle boundary, under the east coast of North America and the Arctic, for example. What is more, the older patches have grown and moved slightly toward the poles. In the late 1980s David Gubbins of the University of Leeds in England—using cruder, older maps of the magnetic field—noticed that the proliferation, growth and poleward migration of these reversed flux patches account for the historical decline of the dipole.

Such observations can be explained physically by using the concept of magnetic lines of force (in actuality, the field is continuous in space). We can think of these lines of force as being "frozen" in the fluid iron core so that they tend to follow its motion, like a filament of dye swirling in a glass of water when stirred. In the earth's core, because of the Coriolis effect, eddies and vortices in the fluid twist magnetic lines of force into bundles that look somewhat like piles of spaghetti. Each twist packs more lines of force into the core, thereby increasing the energy in the magnetic field. (If this process were to go on unchecked, the magnetic field would grow stronger indefinitely. But electrical resistance tends to diffuse and smooth out the twists in the mag-



COMPLEX FLOW PATTERNS in the earth's molten outer core resemble two-dimensional computer simulations of turbulent convection (*left*). When running geodynamo simulations in three dimensions, however, scientists are limited to studying the larger plumes typical of laminar flow (*right*), which is akin to hot mineral oil rising through a lava lamp. Computers are so far incapable of resolving the much more complicated calculations associated with 3-D turbulent flow in the earth's core.

GARY A. GLATZMAIER

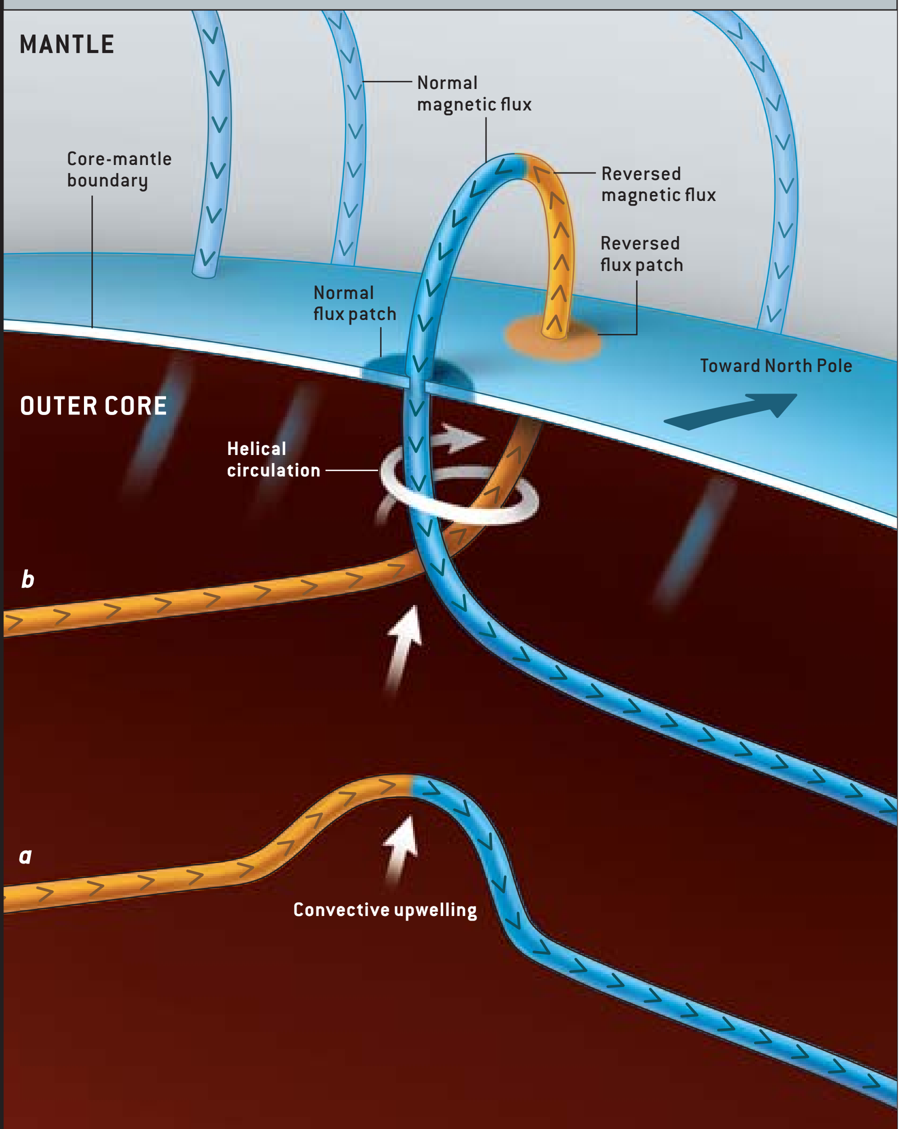
netic field lines enough to suppress runaway growth of the magnetic field without killing the dynamo.)

Patches of intense magnetic flux, both normal and reversed, form on the core-mantle boundary when eddies and vortices interact with east-west-directed magnetic fields, described as toroidal, that are submerged within the core. These turbulent fluid motions can bend and twist the toroidal field lines into loops called poloidal fields, which have a north-south orientation. Sometimes the bending is caused by the rising fluid in an upwelling. If the upwelling is strong enough, the top of the poloidal loop is expelled from the core [*see box on this page*]. This expulsion creates a pair of flux patches where the ends of the loop cross the core-mantle boundary. One of these patches has normally directed flux (in the same direction as the overall dipole field in that hemisphere); the other has the opposite, or reversed, flux.

When the twist causes the reversed flux patch to lie closer to the geographic pole than the normal flux patch, the result is a weakening of the dipole, which is most sensitive to changes near its poles. Indeed, this describes the current situation with the reversed flux patch below the southern tip of Africa. For an actual planetwide polarity reversal to occur, such a reversed flux patch would grow and engulf the entire polar region; at the same time, a similar change in overall regional magnetic polarity would take place near the other geographic pole.

REVERSED FLUX PATCHES

Regions where the direction of magnetic flux is opposite that for the rest of the hemisphere arise when twisted magnetic fields occasionally burst above the earth's core. These reversed flux patches can weaken the main part of the magnetic field at the earth's surface, called the dipole, and may even signal the onset of a global polarity reversal. Reversed flux patches originate as fluid rising through the molten outer core pushes upward on roughly horizontal magnetic field lines within the core. This convective upwelling sometimes bends a line until it bulges (*a*). The earth's rotation simultaneously drives helical circulation of the molten fluid that can twist the bulge into a loop (*b*). When the upwelling force is strong enough to expel the loop from the core, a pair of flux patches forms on the core-mantle boundary.



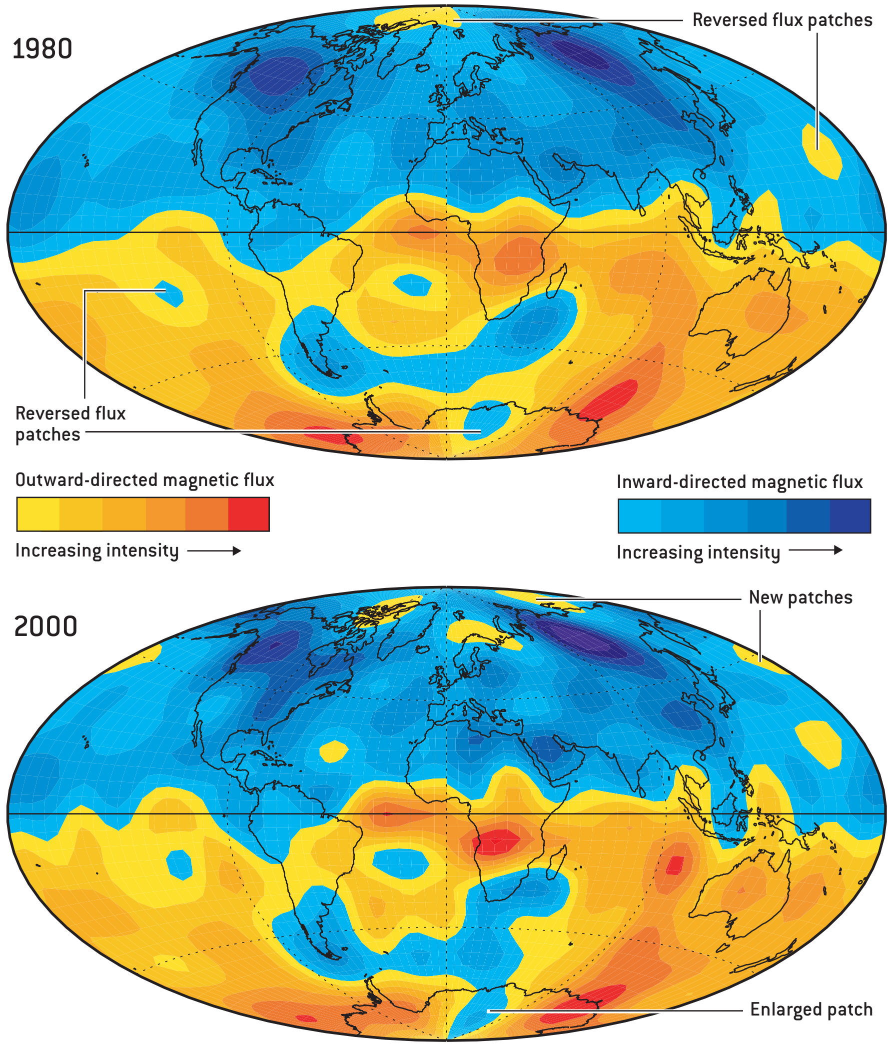
Supercomputer Simulations

TO FURTHER INVESTIGATE how reversed flux patches develop and how they may signal the onset of the next polarity reversal, researchers simulate the geodynamo on supercomputers and in laboratories. The modern era of computer dynamo simulations began in 1995, when three groups—Akira Kageyama of the University of Tokyo and his co-workers; Paul H. Roberts of U.C.L.A. and one of us (Glatzmaier); and Christopher A. Jones of the University of Exeter in England and his colleagues—independently developed numerical simulations that generated magnetic fields resembling the magnetic field at the earth's surface. Since then, simulations representing hundreds of thousands of years have demonstrated how convection can indeed produce patches of reversed magnetic flux on the core-mantle boundary—just like those seen in the satellite images. These patches often appear prior to a spontaneous magnetic dipole reversal, which some simulations can also reproduce.

JEN CHRISTIANSEN

Computer-generated polarity reversals provided researchers with the first rudimentary glimpse of how such switches may originate and progress [*see box on next page*]. One three-dimensional simulation—which had to run for 12 hours a day every day for more than a year to simulate 300,000 years—depicted the onset of a reversal as a decrease in the intensity of the dipole field. Several patches of reversed magnetic flux, such as those now forming on the core-mantle boundary, then began to appear. But rather than extinguishing the magnetic field completely, the reversed flux patches created a weak field with a complex mix of polarities during the transition.

Viewed at the surface of the model earth, the reversal of the dipole occurred when the reversed flux patches begin to dominate the original polarity on the core-mantle boundary. In total, it took about 9,000 years for the old polarity to dissipate and for the new polarity to take hold throughout the core.



CONTOUR MAPS of the earth's magnetic field, extrapolated to the core-mantle boundary from satellite measurements, show that most of the magnetic flux is directed out from the core in the Southern Hemisphere and inward in the Northern Hemisphere. But in a few odd regions, the opposite is true. These so-called reversed flux patches proliferated and grew between 1980 and 2000; if they were to engulf both poles, a polarity reversal could ensue.

PETER OLSON; SOURCE: GAUTHIER HULOT ET AL. IN *NATURE*, VOL. 416, PAGES 620–623; APRIL 11, 2002

What Might Be Missing

BASED IN PART on these successes, computer dynamo models are proliferating rapidly. At last count, more than a dozen groups worldwide were using them to help understand magnetic fields that occur in objects throughout the solar system and beyond. But how well do the geodynamo models capture the dynamo as it actually exists in the earth? The truth is that no one knows for certain.

No computer dynamo model has yet simulated the broad spectrum of turbulence that exists in a planetary interior, primarily because massively parallel supercomputers are not yet fast enough to accurately simulate magnetic turbulence with realistic physical parameters in three dimensions. The smallest turbulent eddies and vortices in the earth's core that twist the magnetic field probably occur on a scale of meters to tens of meters, much less than what can be resolved with the current global geodynamo models on the current supercomputers. That means that all 3-D computer models of the geodynamo so far have simulated the simple, large-scale flow of laminar convection, akin to the hot mineral oil rising through a lava lamp.

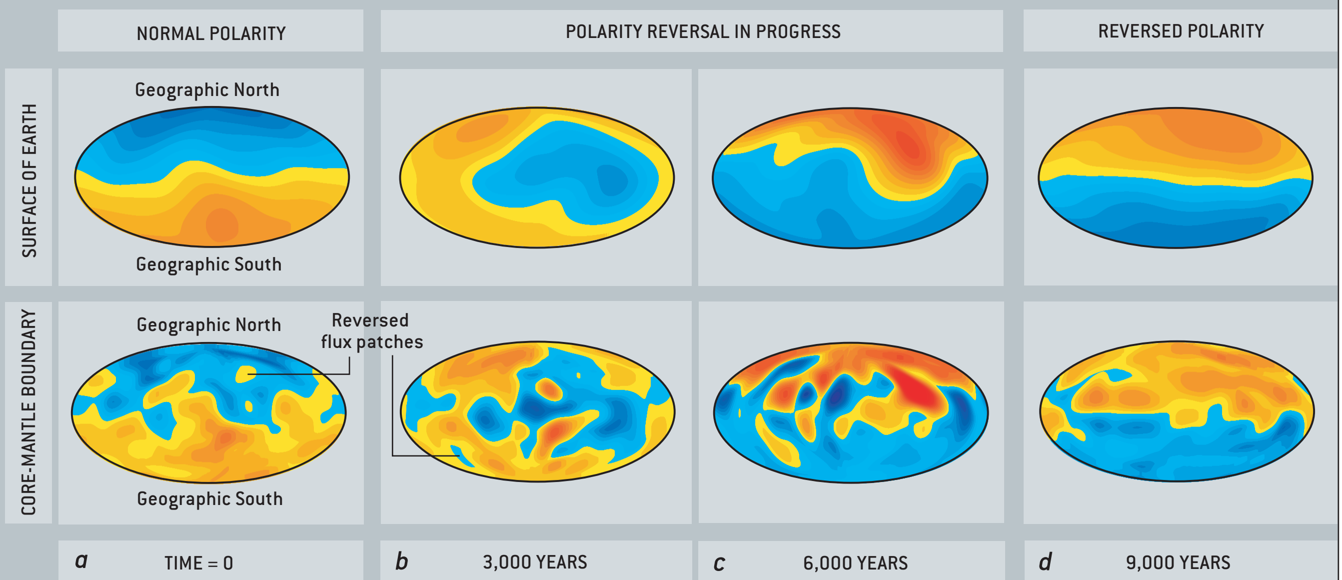
To approximate the effects of turbulent flow in laminar models, investigators can use unrealistically large values for certain properties of the fluid core that, in the real world, are too small to resolve numerically. To achieve realistic turbulence in a computer model, re-

THE AUTHORS

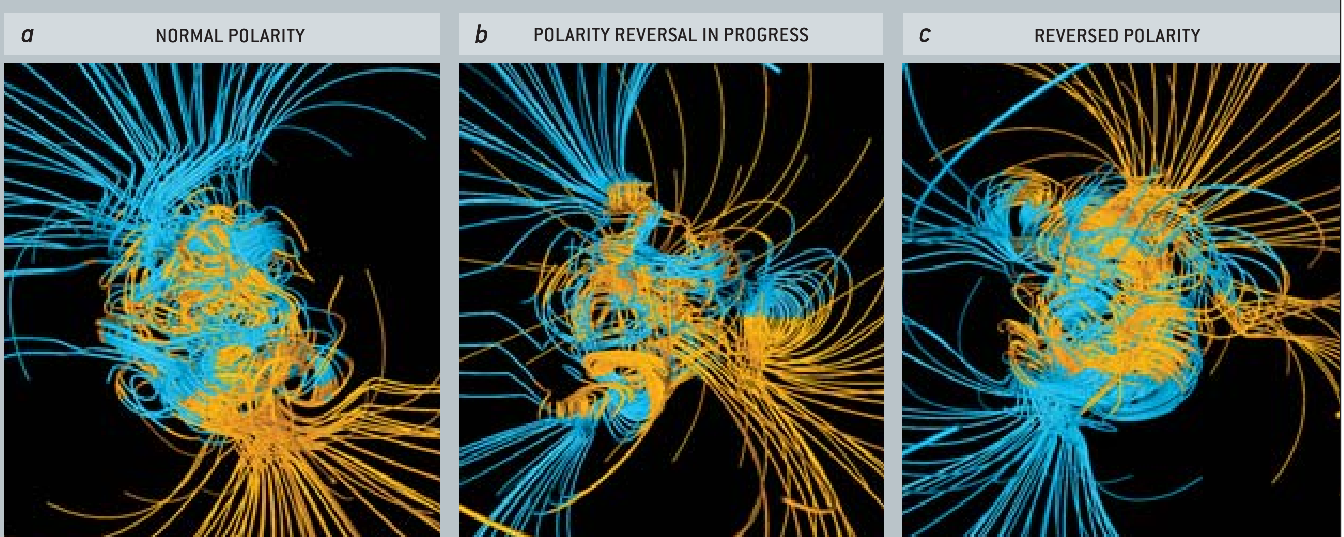
GARY A. GLATZMAIER and *PETER OLSON* develop computer models to study the structure and dynamics of the interiors of planets and stars. In the mid-1990s Glatzmaier, then at Los Alamos National Laboratory (together with Paul H. Roberts of the University of California, Los Angeles), created the first geodynamo simulation that produced a spontaneous magnetic dipole reversal. Glatzmaier has been a professor in the department of earth sciences at the University of California, Santa Cruz, since 1998. Olson is particularly interested in how the earth's core and mantle interact to produce geomagnetic fields, plate tectonics and deep mantle plumes. He joined the department of earth and planetary sciences at Johns Hopkins University in 1978, recently completing a three-year service as departmental chair.

SIMULATED POLARITY REVERSALS

Three-dimensional computer simulations of the geodynamo are now capable of producing spontaneous reversals of the earth's magnetic dipole, offering scientists a way to study the origin of reversals preserved in the geologic record [*see timeline on opposite page*]. One simulated switch typical of a model developed by one of us (Glatzmaier) occurred over a 9,000-year interval. This event is depicted as maps of the vertical part of the magnetic field at the earth's surface and at the core-mantle boundary, where the field is more complex. Models using magnetic field lines provide a third way to visualize a polarity reversal.

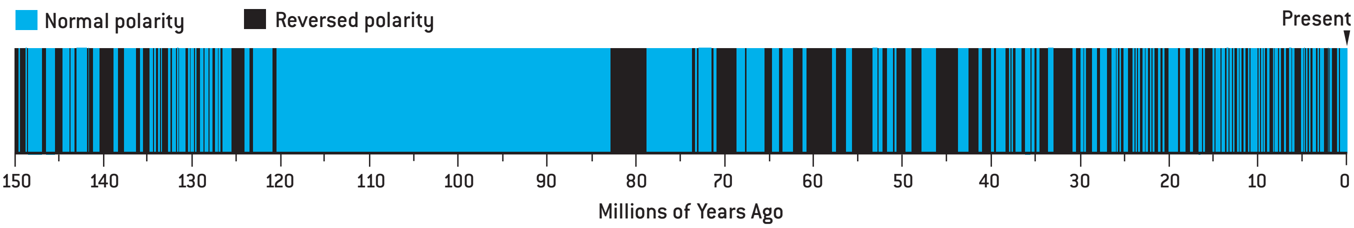


MAGNETIC FIELD MAPS start off with normal polarity, in which most of the overall magnetic flux points out from the core (*yellow*) in the Southern Hemisphere and in toward the core (*blue*) in the Northern Hemisphere (*a*). The onset of the reversal is marked by several areas of reversed magnetic flux (*blue in the south* and *yellow in the north*), reminiscent of the reversed flux patches now forming on the earth's core-mantle boundary. In about 3,000 years the reversed flux patches have decreased the intensity of the dipole field until it is replaced by a weaker but complex transition field at the core-mantle boundary (*b*). The reversal is in full swing by 6,000 years, when the reversed flux patches begin to dominate over the original polarity on the core-mantle boundary (*c*). If viewed only at the surface, the reversal appears complete by this time. But it takes an additional 3,000 years for the dipole to fully reverse throughout the core (*d*).



MODEL illustrates the magnetic field within the core (*tangled lines at center*) and the emerging dipole (*long curved lines*) 500 years before the middle of a magnetic dipole reversal (*a*), at the middle (*b*), and 500 years after that, when the reversal is nearly complete (*c*).

GARY A. GLATZMAIER; SIMULATED AT THE PITTSBURGH SUPERCOMPUTING CENTER



POLARITY REVERSALS have occurred hundreds of times at differing frequencies in the past 150 million years and probably long before that as well. Scientists discovered these reversals by studying magnetic minerals, which line up with the earth's magnetic field when a rock is heated. The minerals then retain the orientation of the field at the time the rock cools.

JEN CHRISTIANSEN; SOURCE: D. V. KENT AND W. LOWRIE IN *TIMESCALES OF THE PALEOMAGNETIC FIELD*. EDITED BY J.E.T. CHANNELL ET AL; AGU, 2004

searchers must resort to a two-dimensional view. The trade-off is that 2-D flow cannot sustain a dynamo. These models do, however, suggest that the laminar flows seen in current geodynamo simulations are much smoother and simpler than the turbulent flows that most likely exist in the earth's core.

Probably the most significant difference is in the paths the fluid follows as it rises through the core. In simple laminar convection simulations, large plumes stretch all the way from the bottom of the core to the top. In the turbulent 2-D models, on the other hand, convection is marked by multiple small-scale plumes and vortices that detach near the upper and lower boundaries of the core and then interact within the main convection zone in-between.

Such differences in the patterns of fluid flow could have a huge influence on the structure of the earth's magnetic field and the time it takes various changes to occur. That is why investigators are diligently pursuing the next generation of 3-D models. Someday, maybe a decade from now, advances in computer processing speeds will make it possible to produce strongly turbulent dynamo simulations. Until then, we hope to learn more from laboratory dynamo experiments now under way.

Laboratory Dynamos

A GOOD WAY to improve understanding of the geodynamo would be to compare computer dynamos (which lack turbulence) with laboratory dynamos (which lack convection). Scientists had first demonstrated the feasibility of lab-scale dynamos in the 1960s, but the road to success was long. The vast difference in size between a laboratory apparatus and the actual core of a planet was a vital factor. A self-sustaining fluid dynamo requires that a certain dimensionless parameter, called the magnetic Reynolds number, exceed a minimum numerical value, roughly 10.

The earth's core has a large magnetic Reynolds number, probably around 1,000, primarily because it has a large linear dimension (the radius of the core is about 3,485 kilometers). Simply put, it is exceedingly difficult to create a large magnetic Reynolds number in small volumes of fluid unless you can move the fluid at extremely high velocities.

The decades-old dream of generating a spontaneous magnetic field in a laboratory fluid dynamo was first realized in 2000, when two groups in Europe—one led by Agris Gailitis of the University of Latvia and one by Robert Stieglitz and Ulrich Müller of the Karlsruhe Research Center and Fritz Busse of the University of Bayreuth, both in Germany—independently achieved self-generation in large volumes of liquid sodium. (Liquid sodium was used because of its high electrical conductivity and low melting point.) Both groups found ways to achieve high-speed fluid flow in a system of one- to two-meter-long helical pipes, resulting in the critical magnetic Reynolds number of about 10.

These experimental results bear out theory, which gives us a measure of confidence when we apply our theoretical ideas about dynamos to the earth and other planets. Now many groups across the world are busy developing the next generation of lab dynamos. To better simulate earthlike geometry, these experiments will stir the liquid sodium inside massive spherical chambers—the largest nearly three meters in diameter.

Besides the ongoing plans for more realistic laboratory dynamos and 3-D computer simulations, the international satellite CHAMP (short for Challenging Minisatellite Payload) is charting the geomagnetic field with enough precision to directly measure its changes at the core-mantle boundary in real time. Investigators anticipate this satellite will provide a continuous image of the geomagnetic field over its five-year mission, allowing them to watch for continued growth of the reversed flux patches as well as other clues about how the dipole field is waning.

We anticipate that a synthesis of these three approaches—satellite observations, computer simulations and laboratory experiments—will occur in the next decade or two. With a more complete picture of the extraordinary geodynamo, we will learn whether our current ideas about the magnetic field and its reversals are on the right track. SA

MORE TO EXPLORE

Numerical Modeling of the Geodynamo: Mechanisms of Field Generation and Equilibration. Peter Olson, Ulrich Christensen and Gary A. Glatzmaier in *Journal of Geophysical Research,* Vol. 104, No. B5, pages 10383–10404; 1999.

Earth's Core and the Geodynamo. Bruce A. Buffett in *Science,* Vol. 288, pages 2007–2012; June 16, 2000.

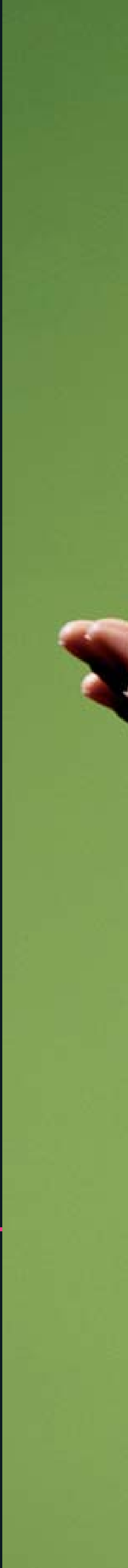
Geodynamo Simulations: How Realistic Are They? Gary A. Glatzmaier in *Annual Review of Earth and Planetary Sciences,* Vol. 30, pages 237–257; 2002.

Recent Geodynamo Simulations and Observations of the Geomagnetic Field. Masaru Kono and Paul H. Roberts in *Reviews of Geophysics,* Vol. 40, No. 4, page 1013; 2002.

The old axiom "one gene, one protein" no longer holds true. The more complex an organism, the more likely it became that way by extracting multiple protein meanings from individual genes

The Alternative Genome

BY GIL AST



GENES of mice and men are 88 percent alike. Many of the ways that humans differ from rodents arise from how we edit our genetic information.

Spring of 2000 found molecular biologists placing dollar bets, trying to predict the number of genes that would be found in the human genome when the sequence of its DNA nucleotides was completed. Estimates at the time ranged as high as 153,000. After all, many said, humans make some 90,000 different types of protein, so we should have at least as many genes to encode them. And given our complexity, we ought to have a bigger genetic assortment than the 1,000-cell roundworm, *Caenorhabditis elegans,* which has a 19,500-gene complement, or corn, with its 40,000 genes.

When a first draft of the human sequence was published the following summer, some observers were therefore shocked by the sequencing team's calculation of 30,000 to 35,000 protein-coding genes. The low number seemed almost embarrassing. In the years since, the human genome map has been finished and the gene estimate has been revised downward still further, to fewer than 25,000. During the same period, however, geneticists have come to understand that our low count might actually be viewed as a mark of our sophistication because humans make such incredibly versatile use of so few genes.

Through a mechanism called alternative splicing, the information stored in the genes of complex organisms can be edited in a variety of ways, making it possible for a single gene to specify two or more distinct proteins. As scientists compare the human genome to those of other organisms, they are realizing the extent to which alternative splicing accounts for much of the diversity among organisms with relatively similar gene sets. In addition, within a single organism, alternative splicing allows different tissue types to perform diverse functions working from the same small gene assortment.

Indeed, the prevalence of alternative splicing appears to increase with an organism's complexity—as many as three quarters of all human genes are subject to alternative editing. The mechanism itself probably contributed to the evolution of that complexity and could drive our further evolution. In the shorter term, scientists are also beginning to understand how faulty gene splicing leads to several cancers and congenital diseases, as well as how the splicing mechanism can be used therapeutically.

Overview/Cut-and-Paste Complexity

- A gene's instructions can be edited by cellular machinery to convey multiple meanings, allowing a small pool of protein-coding genes to give rise to a much larger variety of proteins.
- That such alternative splicing of genetic messages is possible was long understood. But only when the genome sequences of humans and other organisms became available for side-by-side comparison did geneticists see how widespread alternative splicing is in complex organisms and how much the mechanism contributes to differentiating creatures with similar gene sets.
- Alternative splicing enables a minimal number of genes to produce and maintain highly complex organisms by orchestrating when, where and what types of proteins they manufacture. Humans, in turn, may soon be able to regulate our own gene splicing to combat disease.

Pivotal Choices

THE IMPORTANCE of alternative editing to the functioning of many organisms cannot be overestimated. For example, life and death depend on it—at least when a damaged cell must determine whether to go on living. Each cell constantly senses the conditions inside and outside itself, so that it can decide whether to maintain growth or to self-destruct in a preprogrammed process known as apoptosis. Cells that cannot repair DNA will activate their apoptotic program. Craig B. Thompson of the University of Pennsylvania and his colleagues have recently shown that a gene called *Bcl-x,* which is a regulator of apoptosis, is alternatively spliced to produce either of two distinct proteins, Bcl-x(L) and Bcl-x(S). The former suppresses apoptosis, whereas the latter promotes it.

The initial discovery that cells can give rise to such different forms of protein from a single gene was made some 25 years ago, but the phenomenon was considered rare. Recent genome comparisons have revealed it to be both common and crucial, adding a dramatic new twist to the classical view of how information

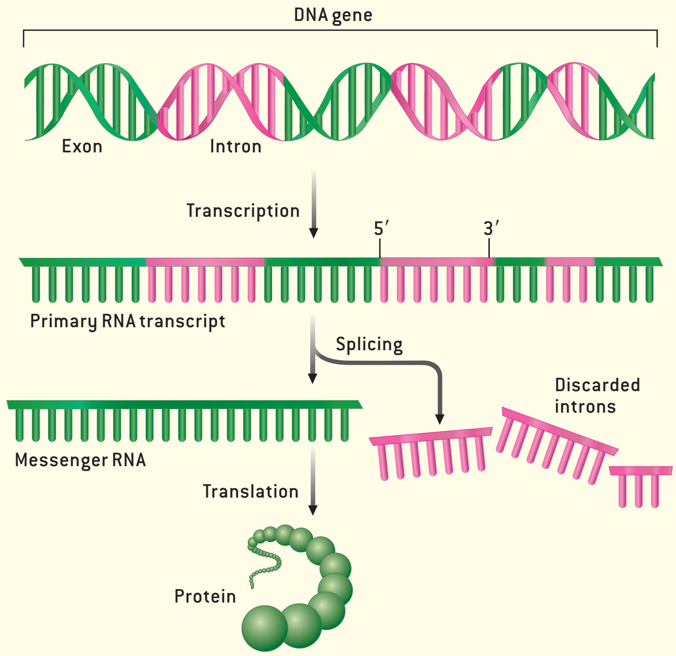
BILL STEELE (*preceding pages*)

ONE GENE, MANY PROTEINS

The classical view of gene expression was simple: a DNA gene is first transcribed into RNA form, then cellular splicing machinery edits out "junk" stretches called introns and joins meaningful portions called exons into a final messenger RNA (mRNA) version, which is then translated into a protein. As it turns out, these rules do not always apply. In complex organisms, the initial RNA transcript can be alternatively spliced—exons may be discarded and introns, or portions of them, retained—to produce a variety of mRNAs, and thus different proteins, from a single gene.

CLASSIC GENE EXPRESSION

A DNA sequence is transcribed into a single-stranded copy made of RNA. Cellular machinery then "splices" this primary transcript: introns—each of which is defined by distinctive nucleotide sequences at its beginning and end, known, respectively, as the 5′ (five-prime) and 3′ (three-prime) splice sites—are removed and discarded while exons are joined into an mRNA version of the gene that will be translated into a protein by the cell.



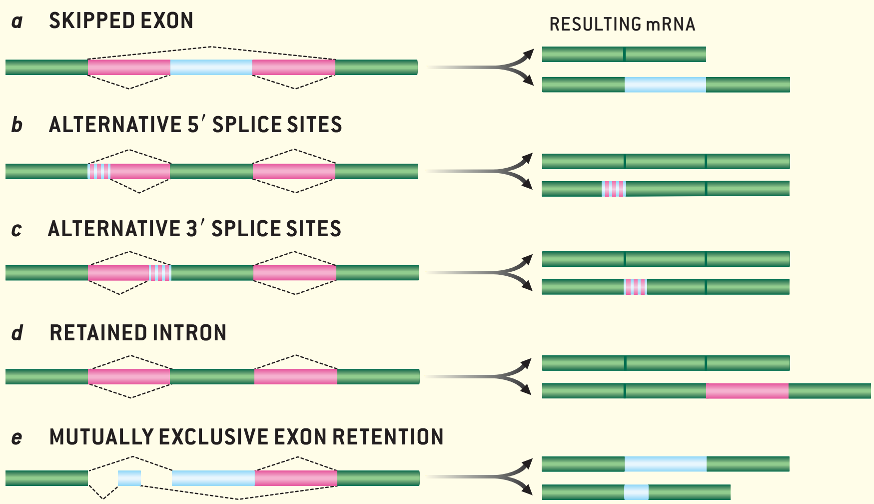
ALTERNATIVE SPLICING

A gene's primary transcript can be edited in several different ways, shown at the right, where splicing activity is indicated by dashed lines. An exon may be left out (*a*). Splicing machinery may recognize alternative 5′ splice sites for an intron (*b*) or alternative 3′ splice sites (*c*). An intron may be retained in the final mRNA transcript (*d*). And exons may be retained on a mutually exclusive basis (*e*).

Exon always spliced in

Exon alternatively spliced

Intron



LUCY READING-IKKANDA

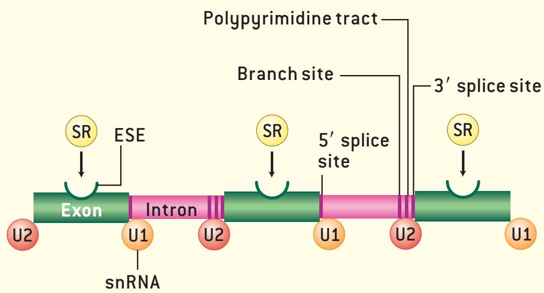
stored in a gene is translated into a protein. Most of the familiar facts still hold true: whole genomes contain all the instructions necessary for making and maintaining an organism, encoded in a four-letter language of DNA nucleotides (abbreviated A, G, C and T). In human chromosomes, roughly three billion nucleotides are strung together on each of two complementary strands that form a double helix. When the time comes for a gene's instructions to be "expressed," the double-stranded zipper of DNA opens just long enough for a single-stranded copy of the gene's sequence to be manufactured from a chemical cousin, RNA. Each sequence of DNA nucleotides that gets transcribed into an RNA version in this manner is called a gene. Some of the resulting RNA molecules are never translated into proteins but rather go on to perform housekeeping and regulatory functions within the cell [see "The Unseen Genome: Gems among the Junk,"

THE SPLICING MACHINE

Once a primary RNA transcript of a gene has been created, a structure called the spliceosome carries out RNA editing. In complex organisms, this process is controlled by splicing regulatory (SR) proteins that define exons and direct the spliceosome to specific splice sites. These regulatory molecules therefore determine when and how alternative splices of a gene will be generated. SR proteins are themselves produced in varying forms in different tissues and cell types or during different stages of development within the same tissue.

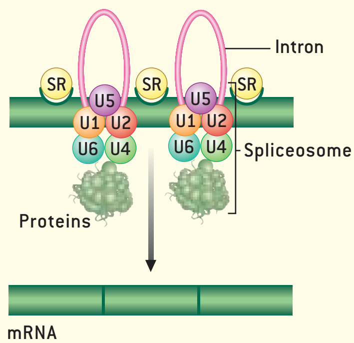
EXON DEFINITION

An SR protein binds to each exon in the transcript at a distinctive nucleotide sequence called an exonic splicing enhancer (ESE). The SR protein's binding defines the exon for the splicing machinery by recruiting small nuclear RNA (snRNA) molecules called U1 and U2 to splice sites on adjacent introns.



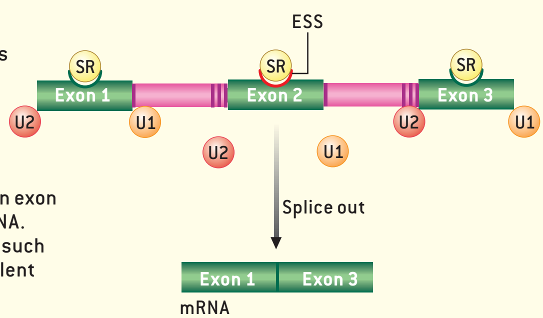
SPLICEOSOME FORMATION

When the original snRNAs have recognized the intron's splice sites, they form a complex with additional snRNAs and more than 100 proteins. This spliceosome complex snips out the introns and joins the exons to produce the mature mRNA.



SPLICING SUPPRESSION

An SR protein may also suppress rather than enhance the binding of snRNAs, in which case the sequence to which it binds is called an exonic splicing suppressor (ESS). The SR protein can thus cause an exon to be spliced out of the final mRNA. In humans and other mammals, such exon skipping is the most prevalent form of alternative splicing.



by W. Wayt Gibbs; SCIENTIFIC AMERICAN, November 2003]. The RNA transcripts of genes that do encode a protein will ultimately be read by cellular machinery and translated into a corresponding sequence of amino acids. But first the preliminary transcript must undergo an editing process.

In 1977 Phillip A. Sharp of the Massachusetts Institute of Technology and Richard J. Roberts of New England Biolabs discovered that these initial, or primary, RNA transcripts are like books containing many nonmeaningful chapters inserted at intervals within the text. The nonsense chapters, called introns, must be excised and the meaningful chapters connected together for the RNA to tell a coherent story. In the cutting-and-ligation process, known as splicing, the introns are snipped out of the primary transcript and discarded. Segments of the transcript containing meaningful protein-coding sequences, called exons, are joined together to form a final version of the transcript, known as messenger RNA (mRNA) [*see box on preceding page*].

But by 1980 Randolph Wall of the University of California at Los Angeles had shown that this basic view of pre-mRNA splicing, in which all introns are always discarded and all exons are always included in the mRNA, does not invariably hold true. In fact, the cellular machinery can "decide" to splice *out* an exon or to leave an intron, or pieces of it, in the final mRNA transcript. This ability to alternatively edit pre-mRNA transcripts can significantly increase any gene's versatility and gives the splicing mechanism tremendous power to determine how much of one type of protein a cell will produce over the other possible types encoded by the same gene.

In 1984 Tom Maniatis, Michael Green and their colleagues at Harvard University developed a test-tube procedure to reveal the molecular machinery that performs the cutting of introns and pasting together of exons. Details of its workings, and of the regulatory system controlling it, are still being filled in, but this research is unveiling an exquisitely intricate system with fascinating origins.

The Splicing Machine

IN COMPLEX ORGANISMS, two distinct levels of molecular equipment are involved in splicing pre-mRNA transcripts. The so-called basal machinery, which is found in all organisms whose genomes contain introns, has been highly conserved through evolutionary time, from yeast to humans. It consists of five small nuclear RNA (snRNA) molecules, identified as U1, U2, U4, U5 and U6. These molecules come together with as many as 150 proteins to form a complex called the spliceosome that is responsible for recognizing the sites where introns begin and end, cutting the introns out of the pre-mRNA transcript and joining the exons to form the mRNA.

Four short nucleotide sequences

LUCY READING-IKKANDA

within introns serve as signals that indicate to the spliceosome where to cut [*see box on opposite page*]. One of these splicing signals sits at the beginning of the intron and is called the 5′ (five-prime) splice site; the others, located at the end of the intron, are known as the branch site, the polypyrimidine tract and, finally, the 3′ (three-prime) splice site.

A separate regulatory system controls the splicing process by directing the basal machinery to these splice sites. More than 10 different splicing regulatory (SR) proteins have been identified. Their forms may vary in different tissues or stages of development in the same tissue. SR proteins can bind to short nucleotide sequences within the exons of the pre-mRNA transcript. These binding sites are known as exonic splicing enhancers (ESE) because when the appropriate SR protein binds to an ESE, that action recruits the basal machinery's snRNAs to the splice sites adjacent to either end of the exon. Yet an SR protein can also bind to an exonic splicing suppressor (ESS) sequence within the exon, which will suppress the basal machinery's ability to bind to the ends of that exon and result in its being spliced out of the final mRNA.

The effect of skipping just one exon can be dramatic for an organism. In fruit flies, for example, alternative splicing regulates the sex-determination pathway. When a gene called *Sex-lethal* is expressed, a male-specific exon may be skipped during splicing, leading to the synthesis of a female-specific Sex-lethal protein. This protein can then bind to any subsequent pre-mRNA transcripts from the same gene, ensuring that all further splicing events will continue to cut out the male-specific exon and guaranteeing that only the female-specific protein will be synthesized. If the male-specific exon is spliced in during the first round of editing, however, a nonfunctional mRNA results, which commits the fly's cells to the male-specific pathway.

The effect of skipping just one exon can be dramatic for an organism.

Exon skipping is the most common type of alternative splicing found in mammals. But several other kinds have also been identified, including one that causes introns to be retained in mature mRNA, which is most prevalent in plants and lower multicellular lifeforms. Intron retention is probably the earliest version of alternative splicing to have evolved. Even today the splicing machinery of single-celled organisms, such as yeast, operates by recognizing introns, in contrast with the SR protein system of higher organisms, which defines exons for the basal machinery.

In the unicellular system the splicing machinery can recognize only intronic sequences of fewer than 500 nucleotides, which works fine for yeast because it has very few introns, averaging just 270 nucleotides long. But as genomes expanded during evolution, their intronic stretches multiplied and grew, and cellular splicing machinery was most likely forced to switch from a system that recognizes short intronic sequences within exons to one that recognizes short exons amid a sea of introns. The average human protein-coding gene, for example, is 28,000 nucleotides long, with 8.8 exons separated by 7.8 introns. The exons are relatively short, usually about 120 nucleotides, whereas the introns can range from 100 to 100,000 nucleotides long.

The size and quantity of human introns—we have the highest number of introns per gene of any organism—raises an interesting issue. Introns are an expensive habit for us to maintain. A large fraction of the energy we consume every day is devoted to the maintenance and repair of introns in their DNA form, transcribing the pre-mRNA and removing the introns, and even to the breakdown of introns at the end of the splicing reaction. Furthermore, this system can cause costly mistakes. Each miscut and ligation of pre-mRNA leads to a change in the gene transcript's protein-coding sequence and possibly to the synthesis of a defective protein.

For instance, an inherited disease that I am investigating, familial dysautonomia, results from a single-nucleotide mutation in a gene called *IKBKAP* that causes it to be alternatively spliced in nervous system tissues. The resulting decreased availability of the standard IKBKAP protein leads to abnormal development of the nervous system, and about half of all patients with this disease die before the age of 30. At least 15 percent of the gene mutations that produce genetic diseases (and probably certain cancers as well) do so by affecting pre-mRNA splicing. So why has evolution preserved such a complicated system that is capable of causing disease? Perhaps because the benefits outweigh the risks.

Advantages in Alternatives

BY GENERATING more than one type of mRNA molecule and, therefore, more than one protein per gene, alternative splicing certainly allows humans to manufacture more than 90,000 proteins without having to maintain 90,000 genes. On average, each of our genes generates about three alternatively spliced mRNAs. Still, that number does not explain our need for so many introns

THE AUTHOR

GIL AST is a senior lecturer in the department of human genetics and molecular medicine at the Tel Aviv University Medical School in Israel. His research focuses on the molecular mechanics of pre-mRNA splicing, the evolution and regulation of alternative splicing, and splicing defects associated with cancers and inherited diseases. He recently collaborated with scientists at Compugen to develop a bioinformatics system for predicting alternative-splicing events to detect novel protein variants.

and why they occupy the vast majority of real estate within genes, leaving exonic sequences to make up only 1 to 2 percent of the human genome.

After the sequencing teams had revealed this seemingly empty genomic landscape in 2001, yet another enigma arose when the mouse genome was published in 2002. It turned out that a mouse possesses almost the same number of genes as a human. Although approximately 100 million years have passed since we had a common ancestor, the vast majority of human and mouse genes derive from that ancestor. Most of these share the same intron and exon arrangement, and the nucleotide sequences within their exons are also conserved to a high degree. So the question becomes, if so little differs between the genomes of humans and mice, what makes us so vastly different from the rodents?

Christopher J. Lee and Barmak Modrek of U.C.L.A. recently revealed that one quarter of the alternatively spliced exons in both genomes are specific either to human or mouse. Thus, these exons have the potential to create species-specific proteins that could be responsible for diversification between species. Indeed, one group of alternatively spliced exons is unique to primates (humans, apes and monkeys) and might have contributed to primates' divergence from other mammals. By studying the process whereby such an exon is born, we can begin to see the advantages of introns in general, and the energy we expend to sustain them seems justified.

These primate-specific exons derive from mobile genetic elements called Alus, which belong to a larger class of elements known as retrotransposons—short sequences of DNA whose function seems to be generating copies of themselves and then reinserting those copies back into the genome at random positions, rather like little genomic parasites. Retrotransposons are found in almost all genomes, and they have had a profound influence by contributing to the genomic expansion that accompanied the evolution of multicellular organisms. Almost half the human genome is made up of transposable elements, Alus being the most abundant.

CHIMPANZEES AND HUMANS share 99 percent of their genomes, including tiny mobile genetic elements, called Alus, found only in primates. Alus may have given rise, through alternative splicing, to new proteins that drove primates' divergence from other mammals. Humans' divergence from other primates may also be thanks in part to alternative splicing: recent studies have shown that the nearly identical genes of humans and chimps produce essentially the same proteins in most tissues, except in parts of the brain, where certain human genes are more active and others generate significantly different proteins through alternative splicing of gene transcripts.

MIKEL ROBERTS *Corbis Sygma*

Alu elements are only 300 nucleotides long with a distinctive sequence that ends in a "poly-A tail." Our genome already contains some 1.4 million Alu copies, and many of these Alu elements are continuing to multiply and insert themselves in new locations in the genome at a rate of about one new insertion per every 100 to 200 human births.

The Alus were long considered nothing more than genomic garbage, but they began to get a little respect as geneticists realized how Alu insertion can expand a gene's protein-generating capacity. About 5 percent of alternatively spliced exons in the human genome contain an Alu sequence. These exons most likely originated when an Alu element "jumped" into an intron of a gene, where the insertion normally would not have any negative consequence for the primate because most introns are spliced out and discarded. Through subsequent mutation, however, the Alu could turn the intron in which it resides into a meaningful sequence of genetic information—an exon. This can happen if changes in the Alu sequence create a new 5′ or 3′ splice site within the intron, causing part of the intron to be recognized as "exon" by the spliceosome. (Such mutations usually arise during cell division, when the genome is copied and a "typo" is introduced.)

If the new Alu exon is only alternatively spliced in, the organism can enjoy the best of two worlds. By including the Alu exon, its cells can produce a novel protein. But the new capability does not interfere with the gene's original function, because the old types of mRNA are also still synthesized when the Alu exon is spliced out. Only when a mutated Alu becomes spliced constitutively—that is, the Alu exon is always spliced into all the mRNAs produced from the gene—does it become problematic, because it can trigger genetic diseases caused by the absence of the old protein. To date, three such genetic illnesses caused by misplaced Alu sequences have been identified: Alport and Sly syndromes and OAT deficiency.

My colleagues and I have shown

that all it takes to convert some silent intronic Alu elements into real exons is a single-letter change in their DNA sequence. At present, the human genome contains approximately 500,000 Alu elements located within introns, and 25,000 of those could become new exons by undergoing this single-point mutation. Thus, Alu sequences have the potential to continue to greatly enrich the stock of meaningful genetic information available for producing new human proteins.

RNA Therapy

MORE THAN 400 research laboratories and some 3,000 scientists worldwide are trying to understand the very complex reactions involved in alternative splicing. Although this research is still at a very early stage, these investigators agree that recent findings point toward future therapeutic applications, such as new gene therapy strategies that exploit the splicing mechanism to treat both inherited and acquired disorders, such as cancer.

One approach might be to direct a short stretch of synthetic RNA or DNA nucleotides, called antisense oligonucleotides, to bind to a specific target on the patient's DNA or RNA. Antisense oligonucleotides could be delivered into cells to mask either a specific splice site or some other regulatory sequence, thereby shifting the splicing activity to another site. Ryszard Kole of the University of North Carolina at Chapel Hill first demonstrated this technique on human blood progenitor cells from patients with an inherited disorder called beta-thalassemia, in which an aberrant 5′ splice site causes oxygen-carrying hemoglobin molecules to be deformed. By masking the mutation, Kole was able to shift splicing back to the normal splice site and restore production of functional hemoglobin.

Later, Kole showed that the same technique could be used on human cancer cells grown in culture. By masking a 5′ splice site of the *Bcl-x* apoptosis-regulating gene transcript, he was able to shift splicing activity to generate the Bcl-x(S) form of mRNA rather than the Bcl-x(L) form, decreasing the cancer cells' synthesis of the antiapoptotic protein and enhancing synthesis of the proapoptotic protein. In some cancer cells, this change activates the apoptotic program; in others, it enhances the apoptotic effects of chemotherapeutic drugs administered along with the oligonucleotides.

Why has evolution preserved a complicated system that can cause disease?

Another way to use the alternative splicing mechanism for therapy was demonstrated in 2003 by Adrian Krainer and Luca Cartegni of Cold Spring Harbor Laboratory in Long Island, N.Y., who found a way to induce cells to splice in an exon that would otherwise be skipped. They created a synthetic molecule that can be programmed to bind to any piece of RNA according to its sequence, then attached the RNA-binding part of an SR protein. This chimeric molecule can therefore both bind to a specified sequence on the pre-mRNA and recruit the basal machinery to the appropriate splice signal. Krainer and Cartegni used this method on human cells grown in culture to correct splicing defects in mutated versions of the *BRCA1* gene, which has been implicated in breast cancer, and of the *SMN2* gene, which causes spinal muscular atrophy.

Yet a third approach capitalizes on the ability of the spliceosome to join two different pre-mRNA molecules from the same gene to form a composite mRNA. Termed trans-splicing, this event is common in worms but occurs only rarely in human cells. Forcing the spliceosome to trans-splice could allow a mutated region of pre-mRNA responsible for disease to be precisely excised and replaced with a normal protein-coding sequence. Recently John Englehardt of the University of Iowa used this technique in cell culture to partially correct the pre-mRNA of a gene that produces a defective protein in the airway cells of cystic fibrosis sufferers.

Before the human genome was decoded, few scientists believed that organisms as complex as humans could survive with a mere 25,000 genes. Since the sequence was completed, alternative splicing has emerged as the pivotal process that permits a small number of genes to generate the much larger assortment of proteins needed to produce the human body and mind while precisely orchestrating their manufacture in different tissues at different times. Moreover, splicing explains how the tremendous diversity among humans, mice and presumably all mammals could originate in such similar genomes.

Evolution works by presenting organisms with new options, then selecting to keep those that confer an advantage. Thus, novel proteins created by the splicing in of new Alu-derived exons probably helped to make humans the species we are today. And further investigation of the alternative splicing process promises still greater improvements in our quality of life. SA

MORE TO EXPLORE

Alternative Splicing: Increasing Diversity in the Proteomic World. B. R. Graveley in *Trends in Genetics,* Vol. 17, No. 2, pages 100–107; February 2001.

Splicing Regulation as a Potential Genetic Modifier. M. Nissim-Rafinia and B. Kerem in *Trends in Genetics,* Vol. 18, No. 3, pages 123–127; March 2002.

The Hidden Genetic Program of Complex Organisms. John S. Mattick in *Scientific American,* Vol. 291, No. 4, pages 60–67; October 2004.

How Did Alternative Splicing Evolve? Gil Ast in *Nature Reviews Genetics,* Vol. 5, pages 773–782; October 2004.

Shaping

By Steven W. Popper, Robert J. Lempert and Steven C. Bankes

the Future

Scientific uncertainty often becomes an excuse to ignore long-term problems, such as climate change. It doesn't have to be so

Last year a high-profile panel of experts known as the Copenhagen Consensus ranked the world's most pressing environmental, health and social problems in a prioritized list. Assembled by the Danish Environmental Assessment Institute under its then director, Bjørn Lomborg, the panel used cost-benefit analysis to evaluate where a limited amount of money would do the most good. It concluded that the highest priority should go to immediate concerns with relatively well understood cures, such as control of malaria. Long-term challenges such as climate change, where the path forward and even the scope of the threat remain unclear, ranked lower.

Usually each of these problems is treated in isolation, as though humanity had the luxury of dealing with its problems one by one. The Copenhagen Consensus used state-of-the-art techniques to try to bring a broader perspective. In so doing, however, it revealed how the state of the art fails to grapple with a simple fact: the future is uncertain. Attempts to predict it have a checkered history—from declarations that humans would never fly, to the doom-and-gloom economic and environmental forecasts of the 1970s, to claims that the "New Economy" would do away with economic ups and downs. Not surprisingly, those who make decisions tend to stay focused on the next fiscal quarter, the next year, the next election. Feeling unsure of their compass, they hug the familiar shore.

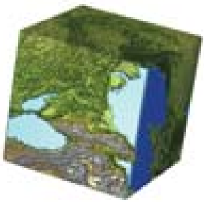
This understandable response to an uncertain future means, however, that the nation's and the world's long-term threats often get ignored altogether or are even made worse by shortsighted decisions. In everyday life, responsible people look out for the long term despite the needs of the here and now: we do homework, we save for retirement, we take out insurance. The same principles should surely apply to society as a whole. But how can leaders weigh the present against the future? How can they avoid being paralyzed by scientific uncertainty?

In well-understood situations, science can reliably predict the implications of alternative policy choices. These predictions, combined with formal methods of decision analysis that use mathematical models and statistical methods to determine optimal courses of action, can specify the trade-offs that society must inevitably make. Corporate executives and elected officials may not always heed this advice, but they do so more often than a cynic might suppose. Analysis has done much to improve the quality of lawmaking, regulation and investment. National economic policy is one example. Concepts introduced by analysts in the 1930s and 1940s—unemployment rate, current-account deficit and gross national product—are now commonplace. For the most part, governments have learned to avoid the radical boom-and-bust cycles that were common in the 19th and early 20th centuries.

The trouble now is that the world faces a number of challenges, both long- and short-term, that are far from well understood: how to preserve the environment, ensure the future of Social Security, guard against terrorism and manage the effects of novel technology. These problems are simply too complex and contingent for scientists to make definitive predictions. In the presence of such deep uncertainty, the machinery of prediction and decision making seizes up. Traditional analytical approaches gravitate to the well-understood parts of the challenge and shy away from the rest. Hence, even sophisticated analyses such as the one by the Copenhagen Consensus have trouble assessing the value of near-term steps that might shape our long-term future.

The three of us—an economist, a physicist and a computer scientist all working in RAND's Pardee Center—have been fundamentally rethinking the role of analysis. We have constructed rigorous, systematic methods for dealing with deep uncertainty. The basic idea is to liberate ourselves from the need for precise prediction by using the computer to help frame strategies that work well over a very wide range of plausible futures. Rather than seeking to eliminate uncertainty, we highlight it and then find ways to manage it. Already companies such as Volvo have used our techniques to plan corporate strategy.

The methods offer a way to break the ideological logjam that too often arises in Washington, D.C. By allowing decision makers to explore a rich variety of what-if scenarios, the new approach reframes the age-old but unanswerable question—What will the long-term future bring?—to one better reflecting our real concern: What actions today will best shape the future to our liking?



The approach replicates the way people reason about UNCERTAIN DECISIONS in everyday life.

The Perils of Prediction

STRIKING A BALANCE between the economy and the environment is one leading example of the difficulty in using science to inform long-term decisions. In his 2002 book *The Future of Life,* Edward O. Wilson described the debate between economists and environmental scientists [see "The Bottleneck," by Edward O. Wilson; SCIENTIFIC AMERICAN, February 2002]. The former group frequently argues that present policies will guide society successfully through the coming century. Technological innovation will reduce pollution and improve energy efficiency, and changes in commodity prices will ensure timely switching from scarce to more plentiful resources. The latter group argues that society's present course will prove unsustainable. By the time the signs of environmental stress become unambiguous, society may have passed the point of easy recovery. Better to apply the brakes now rather than jam them on later when it may be too late.

No matter how compelling their arguments, both sides' detailed predictions are surely wrong. Decisions made today will affect the world 50 to 100 years hence, but no one can credibly predict what life will be like then, regardless of the quality of the science. Interested parties view the same incomplete data, apply different values and assumptions, and arrive at different conclusions. The result can be static and acrimonious debate: "Tree hugger!" "Eco-criminal!"

The (in)famous report *The Limits to Growth* from the early 1970s is the perfect example of how the standard tools of analysis often fail to mediate such debates. A group of scientists and opinion leaders called the Club of Rome predicted that the world would soon exhaust its natural resources unless it took immediate action to slow their use. This conclusion flowed from a then state-of-the-art computer model of the dynamics of resource use. The report met with great skepticism. Since the days of Thomas Malthus, impending resource shortages have melted away as new technologies have made production more efficient and provided alternatives to dwindling resources.

But the model was not wrong; it was just used incorrectly. Any computer model is, by definition, a simplified mirror of the real world, its predictions vulnerable to some neglected factor. The model developed for *The Limits to Growth* revealed some important aspects of the challenges faced by society. In presenting the analysis as a forecast, the authors stretched the model beyond its limits and reduced the credibility of their entire research program.

Overview/*Dealing with Uncertainty*

- Science has become an essential part of decision making by governments and businesses, but uncertainty can foil decision-making frameworks such as cost-benefit analysis. People often end up doing nothing or taking steps that worsen the long-term outlook.
- The authors have developed an alternative framework focused on flexibility—finding, testing and implementing policies that work well no matter what happens.
- Policies can have built-in mechanisms to change with the circumstances. For climate change, one such mechanism is a "safety valve" to ensure that emissions reductions occur but do not get too expensive.

GEORGE RETSECK (*preceding pages*)

Grappling with the Future

CONSCIOUS OF THIS FAILING, analysts have turned to techniques such as scenario planning that involve exploring different possible futures rather than gambling on a single prediction. As an example, in 1995 the Global Scenario Group, convened by the Stockholm Environment Institutes, developed three scenario families. The "Conventional Worlds" family described a future in which technological innovation, driven by markets and lightly steered by government policy, produces economic growth without undermining environmental quality. In the "Barbarization"

PLANNING FOR THE LONG TERM is part of daily life: going to school, taking out insurance and so on. But it is harder to make it part of government and corporate decision making.

FIRST LIGHT *Corbis*

set of scenarios, the same factors—innovation, markets and policy—prove inadequate to the challenge, leading to social collapse and the spread of violence and misery. The third set, "Great Transitions," portrayed the widespread adoption of eco-friendly social values. The Global Scenario Group argued that the Conventional Worlds scenarios are plausible but not guaranteed; to avoid the risk of Barbarization, society should follow the Great Transitions paths.

Although scenario analysis avoids making definite predictions, it has its own shortcomings. It addresses no more than a handful of the many plausible futures, so skeptics can always question the choice of the highlighted few. More fundamentally, scenario families do not translate easily into plans for action. How should decision makers use the scenarios? Should they focus on the most threatening case or the one regarded by experts as most likely? Each approach has faults.

The European Union often favors the "precautionary principle"—in essence, basing policy on the most hazardous plausible scenarios. The Kyoto treaty on climate change, for example, requires reductions of greenhouse gas emissions even though their long-term effects are far from understood. On one level, the precautionary principle makes perfect sense. It is better to be safe than sorry. The long-term future will always be cloudy; some dangers may become certain only when it is too late to prevent them. Yet the principle is an imperfect guide. The future presents many potential harms. Should we worry about them all equally? Few choices are risk-free, and the precautionary principle can lead to contradictory conclusions. For instance, both the harm from greenhouse gas emissions and the cost of reducing them are uncertain. To safeguard the environment, we should reduce the emissions now. To safeguard the economy, we should postpone reductions. So what do we do?

In contrast, many in the U.S. favor cost-benefit analysis, which balances the benefits of eliminating each potential harm against the costs of doing so. When outcomes are uncertain, cost-benefit analysis weights them with odds. We should be willing to pay up to $500 to eliminate a $1,000 harm whose chance of occurring is 50–50. Cost-benefit analysis provides unambiguous answers in many instances. Lead in gasoline enters the environment and affects the developing brains of children. Even though scientists do not know precisely how many children are affected, the benefit of removing lead from gasoline far exceeds the cost. But the long-term future rarely offers such clear choices. Often both the costs and benefits are sufficiently unclear that small disagreements over assigning odds can make a huge difference in the recommended policy.

Making Policies Robust

TRADITIONAL TOOLS such as cost-benefit analysis rely on a "predict then act" paradigm. They require a prediction of the future before they can determine the policy that will work best under the expected circumstances. Because these analyses demand that everyone agree on the models and assumptions, they cannot resolve many of the most crucial debates that our society faces. They force people to select one among many plausible, competing views of the future. Whichever choice emerges is vulnerable to blunders and surprises.

Our approach is to look not for optimal strategies but for robust ones. A robust strategy performs well when compared with the alternatives across a wide range of plausible futures. It need not be the optimal strategy in any future; it will, however, yield satisfactory outcomes in both easy-to-envision futures and hard-to-anticipate contingencies.

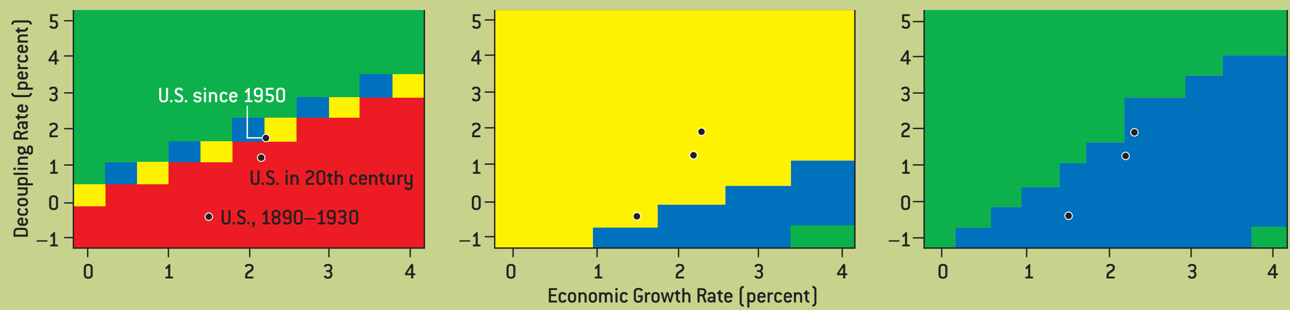
This approach replicates the way people often reason about complicated and uncertain decisions in everyday life. The late Herbert A. Simon, a cognitive scientist and Nobel laureate who pioneered in the 1950s the study of how people make real-world decisions, observed that they seldom optimize. Rather they seek strategies that will work well enough, that include hedges against various potential outcomes and that are

THE AUTHORS

STEVEN W. POPPER, ROBERT J. LEMPERT and *STEVEN C. BANKES* straddle the worlds of science and policymaking. They work at the RAND Corporation in Santa Monica, Calif., one of the country's most renowned think tanks. Popper, an economist, studies how organizations incorporate technological innovation. Lempert, a physicist, specializes in environmental and energy policy. Bankes, a computer scientist, is the father of new methods for computer simulations. All have worked with government and international organizations such as the White House Office of Science and Technology Policy, the Department of Defense, the National Science Foundation and the United Nations. They teach in the Pardee RAND Graduate School and are founders of Evolving Logic, a firm developing software to facilitate the robust decision methods discussed in this article.

BALANCING THE ECONOMY AND THE ENVIRONMENT

How can we clean the planet over the coming century without breaking the bank? The answer depends on how fast the economy will grow and how much existing trends and regulations will cut pollution—and nobody knows either of those things. Many proposed approaches (*left* and *center*) would strike a good balance for some growth rates but not for others, whereas a flexible strategy (*right*) could handle a wide range of scenarios. In the graphs below, colored boxes correspond to particular future rates of growth and "decoupling" (the pace at which existing trends cut pollution). The colors represent how the strategy compares with the theoretically optimal strategy for each scenario: perfectly (■), acceptably (■), poorly (■), or very poorly (■). Dots represent historical rates, which may provide some clue for what is to come.



STAY THE COURSE adds no new environmental policies. It is a good approach if the decoupling rate is high; otherwise, a poor one.

CRASH PROGRAM is an all-out effort to clean up. It justifies itself only if the decoupling rate turns out to be naturally very low.

SAFETY VALVE sets a goal for cutting pollution but relaxes the deadline if the cost runs too high. It works in almost every plausible case.

adaptive. Tomorrow will bring information unavailable today; therefore, people plan on revising their plans.

Incorporating robustness and adaptability into formal decision analysis used to be impossible because of the complexity and vast number of required calculations. Technology has overcome these hurdles. Confronting deep uncertainty requires more than raw computational power, though. The computers have to be used differently. Traditional predict-then-act methods treat the computer as a glorified calculator. Analysts select the model and specify the assumptions; the computer then calculates the optimal strategy implied by these inputs.

In contrast, for robust decision making the computer is integral to the reasoning process. It stress-tests candidate strategies, searching for plausible scenarios that could defeat them. Robust decision making interactively combines the complementary abilities of humans and machines. People excel at seeking patterns, drawing inferences and framing new questions. But they can fail to recognize inconvenient facts and can lose track of how long chains of causes relate to effects. The machine ensures that all claims about strategies are consistent with the data and can reveal scenarios that challenge people's cherished assumptions. No strategy is completely immune to uncertainty, but the computer helps decision makers exploit whatever information they do have to make choices that can endure a wide range of trends and surprises.

Sustainable Development

TO SEE HOW this approach works in practice, return to the dilemma of sustainable development. The first step is to figure out what exactly the computer should calculate. Robust decision making requires the machine to generate multiple paths into the future, spanning the full diversity of those that might occur. We may not know the exact future that will transpire, but any strategy that performs well across a sufficiently diverse set of computer-generated scenarios is likely to meet the challenges presented by what actually comes to pass.

In our analysis of sustainable development, we used a revised version of the Wonderland model originally created by economist Warren C. Sanderson of Stony Brook University and the International Institute for Applied Systems Analysis in Laxenburg, Austria. The Wonderland simulation incorporates, in a very simple manner, scientific understanding of the dynamics of the global economy, demographics and environment. Growing population and wealth will increase pollution, whereas technological innovation may reduce it. The pollution, in turn, hurts the economy when it taxes the environment beyond its absorptive capacity.

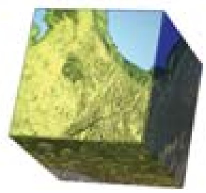
Our version of Wonderland is similar to—but with only 41 uncertain parameters, much simpler than—the simulation used for *The Limits to Growth*. This simplicity can be a virtue: experience demonstrates that additional detail alone does not make predictions more accurate if the model's structure or inputs remain uncertain. For robust planning, models should be used not to predict but to produce a diversity of scenarios, all consistent with the knowledge we do possess.

Running models within special "exploratory modeling" software, analysts can test various strategies and see how they perform. The human user suggests a strategy; for each scenario in the ensemble, the computer compares this approach to the optimal strategy (the one that would have been chosen with perfect predictive foresight) according to such measures as income or life expectancy. A systematic process reveals futures in which the proposed strategies could perform poorly. It also highlights ways each strategy could be adjusted to handle those stressful futures better.

JOHNNY JOHNSON

In the sustainability example, we run the model through the year 2100. Two key uncertainties are the average global economic growth rate during this period and the business-as-usual "decoupling rate" (that is, the reduction in pollution per unit of economic output that would occur in the absence of new environmental policies). The decoupling rate will be positive if existing regulations, productivity increases and the shift to a service economy lessen pollution without lessening growth. It can go negative if growth requires an increase in pollution.

New decision-making methods can break the POLITICAL LOGJAM that arises in Washington.



Depending on the values of these quantities, different strategies perform differently. One strategy, "Stay the Course," simply continues present policy. It performs well in futures where the decoupling rate exceeds the growth rate, but if the reverse is true, pollution eventually becomes so serious that policymakers are forced to abandon the strategy and try to reverse the damage. During the 20th century, the growth and decoupling rates were nearly equal. If the same proves to be true for the 21st, the world will totter on a knife-edge between success and failure [*see box on opposite page*].

The more aggressive "Crash Program" pours money into technological development and environmental regulations that speed decoupling beyond its business-as-usual rate. Although this strategy eliminates the risk of catastrophe, it can impose unnecessarily high costs, inhibiting economic growth.

Becoming Flexible

BOTH THESE STRATEGIES involve policies that are fixed in advance. An adaptive strategy bests them both. Inspired by the complementary strengths and weaknesses of "Stay the Course" and "Crash Program," we considered a flexible alternative that imposes rigorous emissions limits but relaxes them if they cost too much. Such a strategy can be robust. If the technological optimists are right (the decoupling rate turns out to be high), the cost threshold is never breached and industry meets the aggressive environmental goals. If technological pessimists prove correct (the decoupling rate is low), then tight pollution restrictions will exceed the agreed-on cost limits, in which case the strategy gives industry more time to meet the goals.

Such strategies can help cut through contentious debates by providing plans of action that all can agree will play out no matter whose view of the future proves correct. Our adaptive strategy is similar to the "safety valve" strategies that some economists have proposed as alternatives to the immutable emissions targets in the Kyoto treaty. Our new analytical machinery enables decision makers both to design such strategies and to demonstrate their effectiveness to the various interest groups involved.

Of course, even adaptive strategies have their Achilles' heel. In the case of the safety valve, the combination of environmental goals and cost constraints that works best in most futures performs poorly when technological innovation proves to be extremely expensive. To get around this problem, the user can repeat the analysis to come up with a variety of robust strategies, each of which breaks down under different conditions. One strategy may work well when another fails, and vice versa, so the choice between them involves an unavoidable trade-off. The computer calculates how likely each set of circumstances would have to be to justify picking one strategy over the other. Our method thus reduces a complex problem to a small number of simple choices. Decision makers make the final call. Instead of fruitlessly debating models and other assumptions, they can focus on the fundamental trade-offs, fully aware of the surprises that the future may bring.

Clearly, this approach is applicable not only to sustainable development but also to a wide range of other challenges: bringing new products to market, managing the nation's entitlement programs, even defeating terrorism. Science and technology cannot change the future's fundamental unpredictability. Instead they offer an answer to a different question: Which actions today can best usher in a desirable future? Humans and computers search for plausible futures in which a proposed strategy could fail and then identify means to avoid these potential adverse outcomes.

Past failures of prediction should humble anyone who claims to see a clear course into the decades ahead. Paradoxically, though, our greatest possible influence in shaping the future may extend precisely over those timescales where our gaze becomes most dim. We often have little effect on a predictable, near-term future subject to well-understood forces. Where the future is ill defined, unpredictable and hardest to see, our actions today may well have their most profound effects. New tools can help us chart the right course. SA

MORE TO EXPLORE

Exploratory Modeling for Policy Analysis. Steven C. Bankes in *Operations Research*, Vol. 41, No. 3, pages 435–449; May-June 1993.

Assumption-Based Planning. James A. Dewar. Cambridge University Press, 2002.

Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Robert J. Lempert, Steven W. Popper and Steven C. Bankes. RAND MR-1626; 2003. Available online at **www.rand.org/publications/MR/MR1626**

High-Performance Government in an Uncertain World. Robert J. Lempert and Steven W. Popper in *High Performance Government: Structure, Leadership, Incentives.* Edited by Robert Klitgaard and Paul C. Light. RAND-MG-256; 2005.

Humans and other animals share a heritage of economic tendencies—including cooperation, repayment of favors and resentment at being shortchanged

How Animals Do Business

By Frans B. M. de Waal

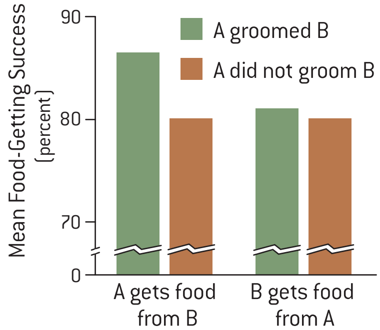
Just as my office would not stay empty for long were I to move out, nature's real estate changes hands all the time. Potential homes range from holes drilled by woodpeckers to empty shells on the beach. A typical example of what economists call a "vacancy chain" is the housing market among hermit crabs. To protect its soft abdomen, each crab carries its house around, usually an abandoned gastropod shell. The problem is that the crab grows, whereas its house does not. Hermit crabs are always on the lookout for new accommodations. The moment they upgrade to a roomier shell, other crabs line up for the vacated one.

One can easily see supply and demand at work here, but because it plays itself out on a rather impersonal level, few would view the crab version as related to human economic transactions. The crab interactions would be more interesting if the animals struck deals along the lines of "you can have my house if I can have that dead fish." Hermit crabs are not deal makers, though, and in fact have no qualms about evicting homeowners by force. Other, more social animals do negotiate, however, and their approach to the exchange of resources and services helps us understand how and why human economic behavior may have evolved.

The New Economics

CLASSICAL ECONOMICS views people as profit maximizers driven by pure selfishness. As 17th-century English philosopher Thomas Hobbes put it, "Every man is presumed to seek what is good for himselfe naturally, and what is just, only for Peaces sake, and accidentally." In this still prevailing view, sociality is but an afterthought, a "social contract" that our ancestors en-

CAPUCHIN MONKEYS share food just as chimpanzees and humans do. Rare among other primates, this practice may have evolved along with cooperative hunting, a strategy used by all three species. Without joint payoffs, there would be no joint hunting. Here a juvenile capuchin begs for a share by cupping his hand next to the food an adult male is eating.



CHIMPANZEES share food—these branches with leaves, for example—in return for favors such as grooming. This reciprocity was demonstrated experimentally by recording groomings on the mornings of days when food-sharing tests were scheduled. As the graph shows, chimp A's success in obtaining food from chimp B increased after A had groomed B, but B's success in obtaining food from A was unaffected by A's grooming. Thus, it is specifically the groomer who benefits, meaning that the rule is one of exchange of food for grooming.

tered into because of its benefits, not because they were attracted to one another. For the biologist, this imaginary history falls as wide off the mark as can be. We descend from a long line of group-living primates, meaning that we are naturally equipped with a strong desire to fit in and find partners to live and work with. This evolutionary explanation for why we interact as we do is gaining influence with the advent of a new school, known as behavioral economics, that focuses on actual human behavior rather than on the abstract forces of the marketplace as a guide for understanding economic decision making. In 2002 the school was recognized by a shared Nobel Prize for two of its founders: Daniel Kahneman and Vernon L. Smith.

Animal behavioral economics is a fledgling field that lends support to the new theories by showing that basic human economic tendencies and preoccupations—such as reciprocity, the division of rewards, and cooperation—are not limited to our species. They probably evolved in other animals for the same reasons they evolved in us—to help individuals take optimal advantage of one another without undermining the shared interests that support group life.

Take a recent incident during my research at the Yerkes National Primate Research Center in Atlanta. We had taught capuchin monkeys to reach a cup of food on a tray by pulling on a bar attached to the tray. By making the tray too heavy for a single individual, we gave the monkeys a reason to work together.

On one occasion, the pulling was to be done by two females, Bias and Sammy. Sitting in adjoining cages, they successfully brought a tray bearing two cups of food within reach. Sammy, however, was in such a hurry to collect her reward that she released the bar and grabbed her cup before Bias had a chance to get hers. The tray bounced back, out of Bias's reach. While Sammy munched away, Bias threw a tantrum. She screamed her lungs out for half a minute until Sammy approached her pull bar again. She then helped Bias bring in the tray a second time. Sammy did not do so for her own benefit, because by now the cup accessible to her was empty.

Sammy's corrective behavior appeared to be a response to Bias's protest against the loss of an anticipated reward. Such action comes much closer to human economic transactions than that of the hermit crabs, because it shows cooperation, communication and the fulfillment of an expectation, perhaps even a sense of obligation. Sammy seemed sensitive to the quid pro quo of the situation. This sensitivity is not surprising given that the group life of capuchin monkeys revolves around the same mixture of cooperation and competition that marks our own societies.

Overview/*Evolved Economics*

- The new field of behavioral economics views the way humans conduct business as an evolved heritage of our species.
- Just as tit for tat and supply and demand influence the trading of goods and services in human economies, they also affect trading activities among animals.
- Emotional reactions—such as outrage at unfair arrangements—underlie the negotiations of both animals and humans.
- This shared psychology may explain such curious behaviors as altruism—they are part of our background as cooperative primates.

The Evolution of Reciprocity

ANIMALS AND PEOPLE occasionally help one another without any obvious benefits for the helper. How could such behavior have evolved? If the aid is di-

FRANS B. M. DE WAAL (*preceding pages and this page*); ALISON KENDALL (*graph*) rected at a family member, the question is relatively easy to answer. "Blood is thicker than water," we say, and biologists recognize genetic advantages to such assistance: if your kin survive, the odds of your genes making their way into the next generation increase. But cooperation among unrelated individuals suggests no immediate genetic advantages. Pëtr Kropotkin, a Russian prince, offered an early explanation in his book *Mutual Aid,* published in 1902. If helping is communal, he reasoned, all parties stand to gain—everyone's chances for survival go up. We had to wait until 1971, however, for Robert L. Trivers, then at Harvard University, to phrase the issue in modern evolutionary terms with his theory of reciprocal altruism.

Trivers contended that making a sacrifice for another pays off if the other later returns the favor. Reciprocity boils down to "I'll scratch your back, if you scratch mine." Do animals show such tit for tat? Monkeys and apes form coalitions; two or more individuals, for example, gang up on a third. And researchers have found a positive correlation between how often A supports B and how often B supports A. But does this mean that animals actually keep track of given and received favors? They may just divide the world into "buddies," whom they prefer, and "nonbuddies," whom they care little about. If such feelings are mutual, relationships will be either mutually helpful or mutually unhelpful. Such symmetries can account for the reciprocity reported for fish, vampire bats (which regurgitate blood to their buddies), dolphins and many monkeys.

Just because these animals may not keep track of favors does not mean they lack reciprocity. The issue rather is how a favor done for another finds its way back to the original altruist. What exactly is the reciprocity mechanism? Mental record keeping is just one way of getting reciprocity to work, and whether animals do this remains to be tested. Thus far chimpanzees are the only exception. In the wild, they hunt in teams to capture colobus monkeys. One hunter usually captures the prey, after which he tears it apart and shares it. Not everyone gets a piece, though, and even the highest-ranking male, if he did not take part in the hunt, may beg in vain. This by itself suggests reciprocity: hunters seem to enjoy priority during the division of spoils.

To try to find the mechanisms at work here, we exploited the tendency of these apes to share—which they also show in captivity—by handing one of the chimpanzees in our colony a watermelon or some branches with leaves. The owner would be at the center of a sharing cluster, soon to be followed by secondary clusters around individuals who had managed to get a major share, until all the food had trickled down to everyone. Claiming another's food by force is almost unheard of among chimpanzees—a phenomenon known as "re-

What Makes Reciprocity Tick

Humans and other animals exchange benefits in several ways, known technically as reciprocity mechanisms. No matter what the mechanism, the common thread is that benefits find their way back to the original giver.

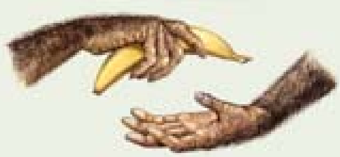
RECIPROCITY MECHANISM — **KEY FEATURES**

Symmetry-based
"We're buddies"

Mutual affection between two parties prompts similar behavior in both directions without need to keep track of daily give-and-take, so long as the overall relationship remains satisfactory. Possibly the most common mechanism of reciprocity in nature, this kind is typical of humans and chimpanzees in close relationships.

Example: Chimpanzee friends associate, groom together and support each other in fights.

Attitudinal
"If you're nice, I'll be nice"



Parties mirror one another's attitudes, exchanging favors on the spot. Instant attitudinal reciprocity occurs among monkeys, and people often rely on it with strangers.

Example: Capuchins share food with those who help them pull a treat-laden tray.

Calculated
"What have you done for me lately?"

Individuals keep track of the benefits they exchange with particular partners, which helps them decide to whom to return favors. This mechanism is typical of chimpanzees and common among people in distant and professional relationships.

Example: Chimpanzees can expect food in the afternoon from those they groomed in the morning.

ROBERTO OSTI

CLEANER FISH nibbles parasites in the open mouth of a large client fish. Roaming client fish rarely return to the station of a cleaner fish after they have been kept waiting (*left graph*) or cheated (*right graph*), meaning that the cleaner took a bite out of the client's healthy tissue. Cleaner fish therefore tend to treat roaming clients better than residents, who have no choice of cleaning stations.

spect of possession." Beggars hold out their hand, palm upward, much like human beggars in the street. They whimper and whine, but aggressive confrontations are rare. If these do occur, the possessor almost always initiates them to make someone leave the circle. She whacks the offenders over the head with a sizable branch or barks at them in a shrill voice until they leave her alone. Whatever their rank, possessors control the food flow.

We analyzed nearly 7,000 of these approaches, comparing the possessor's tolerance of specific beggars with previously received services. We had detailed records of grooming on the mornings of days with planned food tests. If the top male, Socko, had groomed May, for example, his chances of obtaining a few branches from her in the afternoon were much improved. This relation between past and present behavior proved general. Symmetrical connections could not explain this outcome, as the pattern varied from day to day. Ours was the first animal study to demonstrate a contingency between favors given and received. Moreover, these food-for-grooming deals were partner-specific—that is, May's tolerance benefited Socko, the one who had groomed her, but no one else.

This reciprocity mechanism requires memory of previous events as well as the coloring of memory such that it induces friendly behavior. In our own species, this coloring process is known as "gratitude," and there is no reason to call it something else in chimpanzees. Whether apes also feel obligations remains unclear, but it is interesting that the tendency to return favors is not the same for all relationships. Between individuals who associate and groom a great deal, a single grooming session carries little weight. All kinds of daily exchanges occur between them, probably without their keeping track. They seem instead to follow the buddy system discussed before. Only in the more distant relationships does grooming stand out as specifically deserving reward. Because Socko and May are not close friends, Socko's grooming was duly noticed.

A similar difference is apparent in human behavior, where we are more inclined to keep track of give-and-take with strangers and colleagues than with our friends and family. In fact, scorekeeping in close relationships, such as between spouses, is a sure sign of distrust.

Biological Markets

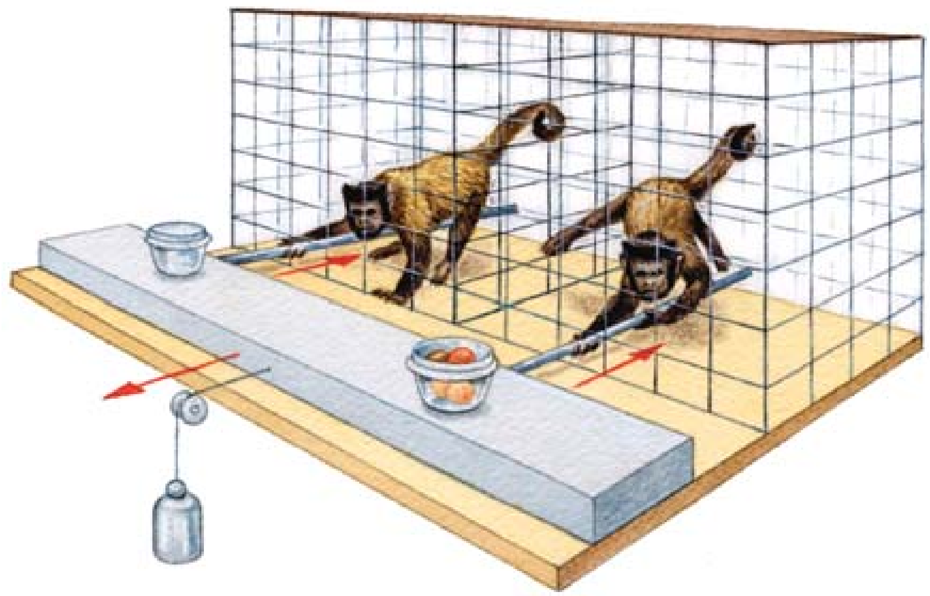
BECAUSE RECIPROCITY requires partners, partner choice ranks as a central issue in behavioral economics. The

BABOON FEMALES pay a price in grooming to get a peek at a new infant; the fewer the infants, the longer the grooming time required. The value of commodities—baby baboons in this case—increases as their availability decreases.

TIM LAMAN *National Geographic Image Collection* (*top*); S. P. HENZI (*bottom*); ALISON KENDALL (*graph*) hand-me-down housing of hermit crabs is exceedingly simple compared with the interactions among primates, which involve multiple partners exchanging multiple currencies, such as grooming, sex, support in fights, food, babysitting and so on. This "marketplace of services," as I dubbed it in *Chimpanzee Politics*, means that each individual needs to be on good terms with higher-ups, to foster grooming partnerships and—if ambitious—to strike deals with like-minded others. Chimpanzee males form coalitions to challenge the reigning ruler, a process fraught with risk. After an overthrow, the new ruler needs to keep his supporters contented: an alpha male who tries to monopolize the privileges of power, such as access to females, is unlikely to keep his position for long. And chimps do this without having read Niccolò Machiavelli.

With each individual shopping for the best partners and selling its own services, the framework for reciprocity becomes one of supply and demand, which is precisely what Ronald Noë and Peter Hammerstein, then at the Max Planck Institute for Behavioral Physiology in Seewiesen, Germany, had in mind with their biological market theory. This theory, which applies whenever trading partners can choose with whom to deal, postulates that the value of commodities and partners varies with their availability. Two studies of market forces elaborate this point: one concerns the baby market among baboons, the other the job performance of small fish called cleaner wrasses.

Like all primate females, female baboons are irresistibly attracted to infants—not only their own but also those of others. They give friendly grunts and try to touch them. Mothers are highly protective, however, and reluctant to let anyone handle their precious newborns. To get close, interested females groom the mother while peeking over her shoulder or underneath her arm at the baby. After a relaxing grooming session, a mother may give in to the groomer's desire for a closer look. The other thus buys infant time. Market theory predicts that the value of babies should go up if there are fewer around. In a study of wild chacma baboons in South Africa, Louise Barrett of the University of Liverpool and Peter Henzi of the University of Central Lancashire, both in England, found that, indeed, mothers of rare infants were able to extract a higher price (longer grooming) than mothers in a troop full of babies.



TRAY-PULLING EXPERIMENT demonstrates that capuchin monkeys are more likely to share food with cooperative partners than with those who are not helpful. The test chamber houses two capuchins, separated by mesh. To reach their treat cups, they must use a bar to pull a counterweighted tray; the tray is too heavy for one monkey to handle alone. The "laborer" (*on left*), whose transparent cup is obviously empty, works for the "winner," who has food in its cup. The winner generally shares food with the laborer through the mesh. Failing to do so will cause the laborer to lose interest in the task.

ROBERTO OSTI

Cleaner wrasses (*Labroides dimidiatus*) are small marine fish that feed on the external parasites of larger fish. Each cleaner owns a "station" on a reef where clientele come to spread their pectoral fins and adopt postures that offer the cleaner a chance to do its job. The exchange exemplifies a perfect mutualism.

The cleaner nibbles the parasites off the client's body surface, gills and even the inside of its mouth. Sometimes the cleaner is so busy that clients have to wait in line. Client fish come in two varieties: residents and roamers. Residents belong to species with small territories; they have no choice but to go to their local cleaner. Roamers, on the other hand, either hold large territories or travel widely, which means that they have several cleaning stations to choose from. They want short waiting times, excellent service and no cheating. Cheating occurs when a cleaner fish takes a bite out of its client, feeding on healthy mucus. This makes clients jolt and swim away.

Research on cleaner wrasses by Redouan Bshary of the Max Planck institute in Seewiesen consists mainly of observations on the reef but also includes ingenious experiments in the laboratory. His papers read much like a manual for good business practice. Roamers are more likely to change stations if a cleaner has ignored them for too long or cheated them. Cleaners seem to know this and treat roamers better than they do residents. If a roamer and a resident arrive at the same time, the cleaner almost always services the roamer first. Residents

THE AUTHOR

FRANS B. M. DE WAAL is C. H. Candler Professor of Primate Behavior at Emory University and director of the Living Links Center at the university's Yerkes National Primate Research Center. De Waal specializes in the social behavior and cognition of monkeys, chimpanzees and bonobos, especially cooperation, conflict resolution and culture. His books include *Chimpanzee Politics, Peacemaking among Primates, The Ape and the Sushi Master* and the forthcoming *Our Inner Ape*.

have nowhere else to go, and so they can be kept waiting. The only category of fish that cleaners never cheat are predators, who possess a radical counterstrategy, which is to swallow the cleaner. With predators, cleaner fish wisely adopt, in Bshary's words, an "unconditionally cooperative strategy."

Biological market theory offers an elegant solution to the problem of free-loaders, which has occupied biologists for a long time because reciprocity systems are obviously vulnerable to those who take rather than give. Theorists often assume that offenders must be punished, although this has yet to be demonstrated for animals. Instead cheaters can be taken care of in a much simpler way. If there is a choice of partners, animals can simply abandon unsatisfactory relationships and replace them with those offering more benefits. Market mechanisms are all that is needed to sideline profiteers. In our own societies, too, we neither like nor trust those who take more than they give, and we tend to stay away from them.

Fair Is Fair

TO REAP THE BENEFITS of cooperation, an individual must monitor its efforts relative to others and compare its rewards with the effort put in. To explore whether animals actually carry out such monitoring, we turned again to our capuchin monkeys, testing them in a miniature labor market inspired by field observations of capuchins attacking giant squirrels. Squirrel hunting is a group effort, but one in which all rewards end up in the hands of a single individual: the captor. If captors were to keep the prey solely for themselves, one can imagine that others would lose interest in joining them in the future. Capuchins share meat for the same reason chimpanzees (and people) do: there can be no joint hunting without joint payoffs.

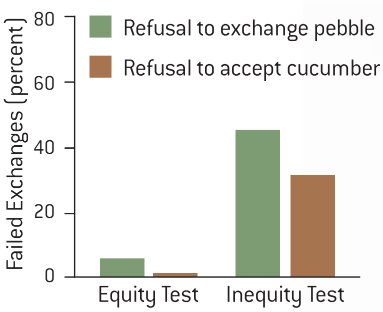
We mimicked this situation in the laboratory by making certain that only one monkey (whom we called the winner) of a tray-pulling pair received a cup with apple pieces. Its partner (the laborer) had no food in its cup, which was obvious from the outset because the cups were transparent. Hence, the laborer pulled for the winner's benefit. The monkeys sat side by side, separated by mesh. From previous tests we knew that food possessors might bring food to the partition and permit their neighbor to reach for it through the mesh. On rare occasions, they push pieces to the other.

We contrasted collective pulls with solo pulls. In one condition, both animals had a pull bar and the tray was heavy; in the other, the partner lacked a bar and the winner handled a lighter tray on its own. We counted more acts of food sharing after collective than solo pulls: winners were in effect compensating their partners for the assistance they had received. We also confirmed that sharing affects future cooperation. Because a pair's success rate would drop if the winner failed to share, payment of the laborer was a smart strategy.

Sarah F. Brosnan, one of my colleagues at Yerkes, went further in exploring reactions to the way rewards are divided. She would offer a capuchin monkey a small pebble, then hold up a slice of cucumber as enticement for returning the pebble. The monkeys quickly grasped the principle of exchange. Placed side by side, two monkeys would gladly exchange pebbles for cucumber with the researcher. If one of them got grapes, however, whereas the other stayed on cucumber, things took an unexpected turn. Grapes are much preferred. Monkeys who had been perfectly willing to work for cucumber suddenly went on strike. Not only did they perform reluctantly seeing that the other was getting a better deal, but they became agitated, hurling the pebbles out of the test chamber and sometimes even the cucumber slices. A food normally never refused had become less than desirable.

To reject unequal pay—which people do as well—goes against the assump-

FRANS B. M. DE WAAL; ALISON KENDALL (*graph*)



CAPUCHIN MONKEYS have definite preferences when it comes to food. They will, for example, choose fruit over vegetables, such as the celery this capuchin is thoughtfully consuming. Trained to exchange a pebble for a slice of cucumber, they happily did so as long as the monkey in the adjoining test chamber also received cucumber (*Equity Test on graph*). But when the monkey next door was given a grape while they continued to receive cucumber (*Inequity Test*), they balked at "unfair pay." They either refused to accept the cucumber, sometimes even throwing it out of the cage, or refused to return the pebble.

How Humans Do Business

The emotions that Frans de Waal describes in the economic exchanges of social animals have parallels in our own transactions. Such similarities suggest that human economic interactions are controlled at least in part by ancient tendencies and emotions. Indeed, the animal work supports a burgeoning school of research known as behavioral economics. This new discipline is challenging and modifying the "standard model" of economic research, which maintains that humans base economic decisions on rational thought processes. For example, people reject offers that strike them as unfair, whereas classical economics predicts that people take anything they can get. In 2002 the Nobel Prize in Economics went to two pioneers of the field: Daniel Kahneman, a psychologist at Princeton University, and Vernon L. Smith, an economist at George Mason University.

Kahneman, with his colleague Amos Tversky, who died in 1996 and thus was not eligible for the prize, analyzed how humans make decisions when confronted by uncertainty and risk. Classical economists had thought of human decisions in terms of expected utility—the sum of the gains people think they will get from some future event multiplied by its probability of occurring. But Kahneman and Tversky demonstrated that people are much more frightened of losses than they are encouraged by potential gains and that people follow the herd. The bursting of the stock-market bubble in 2000 provides a potent example: the desire to stay with the herd may have led people to shell out far more for shares than any purely rational investor would have paid.


IRRATIONAL EXUBERANCE can grip the floor of a stock exchange.

Smith's work demonstrated that laboratory experiments would function in economics, which had traditionally been considered a nonexperimental science that relied solely on observation. Among his findings in the lab: emotional decisions are not necessarily unwise. *—The Editors*

tions of traditional economics. If maximizing benefits were all that mattered, one should take what one can get and never let resentment or envy interfere. Behavioral economists, on the other hand, assume evolution has led to emotions that preserve the spirit of cooperation and that such emotions powerfully influence behavior. In the short run, caring about what others get may seem irrational, but in the long run it keeps one from being taken advantage of. Discouraging exploitation is critical for continued cooperation.

It is a lot of trouble, though, to always keep a watchful eye on the flow of benefits and favors. This is why humans protect themselves against freeloading and exploitation by forming buddy relationships with partners—such as spouses and good friends—who have withstood the test of time. Once we have determined whom to trust, we relax the rules. Only with more distant partners do we keep mental records and react strongly to imbalances, calling them "unfair."

We found indications for the same effect of social distance in chimpanzees. Straight tit for tat, as we have seen, is rare among friends who routinely do favors for one another. These relationships also seem relatively immune to inequity. Brosnan conducted her exchange task using grapes and cucumbers with chimpanzees as well as capuchins. The strongest reaction among chimpanzees concerned those who had known one another for a relatively short time, whereas the members of a colony that had lived together for more than 30 years hardly reacted at all. Possibly, the greater their familiarity, the longer the time frame over which chimpanzees evaluate their relationships. Only distant relations are sensitive to day-to-day fluctuations.

All economic agents, whether human or animal, need to come to grips with the freeloader problem and the way yields are divided after joint efforts. They do so by sharing most with those who help them most and by displaying strong emotional reactions to violated expectations. A truly evolutionary discipline of economics recognizes this shared psychology and considers the possibility that we embrace the golden rule not accidentally, as Hobbes thought, but as part of our background as cooperative primates. SA

MORE TO EXPLORE

The Chimpanzee's Service Economy: Food for Grooming. Frans B. M. de Waal in *Evolution and Human Behavior,* Vol. 18, No. 6, pages 375–386; November 1997.

Payment for Labour in Monkeys. Frans B. M. de Waal and Michelle L. Berger in *Nature,* Vol. 404, page 563; April 6, 2000.

Choosy Reef Fish Select Cleaner Fish That Provide High-Quality Service. R. Bshary and D. Schäffer in *Animal Behaviour,* Vol. 63, No. 3, pages 557–564; March 2002.

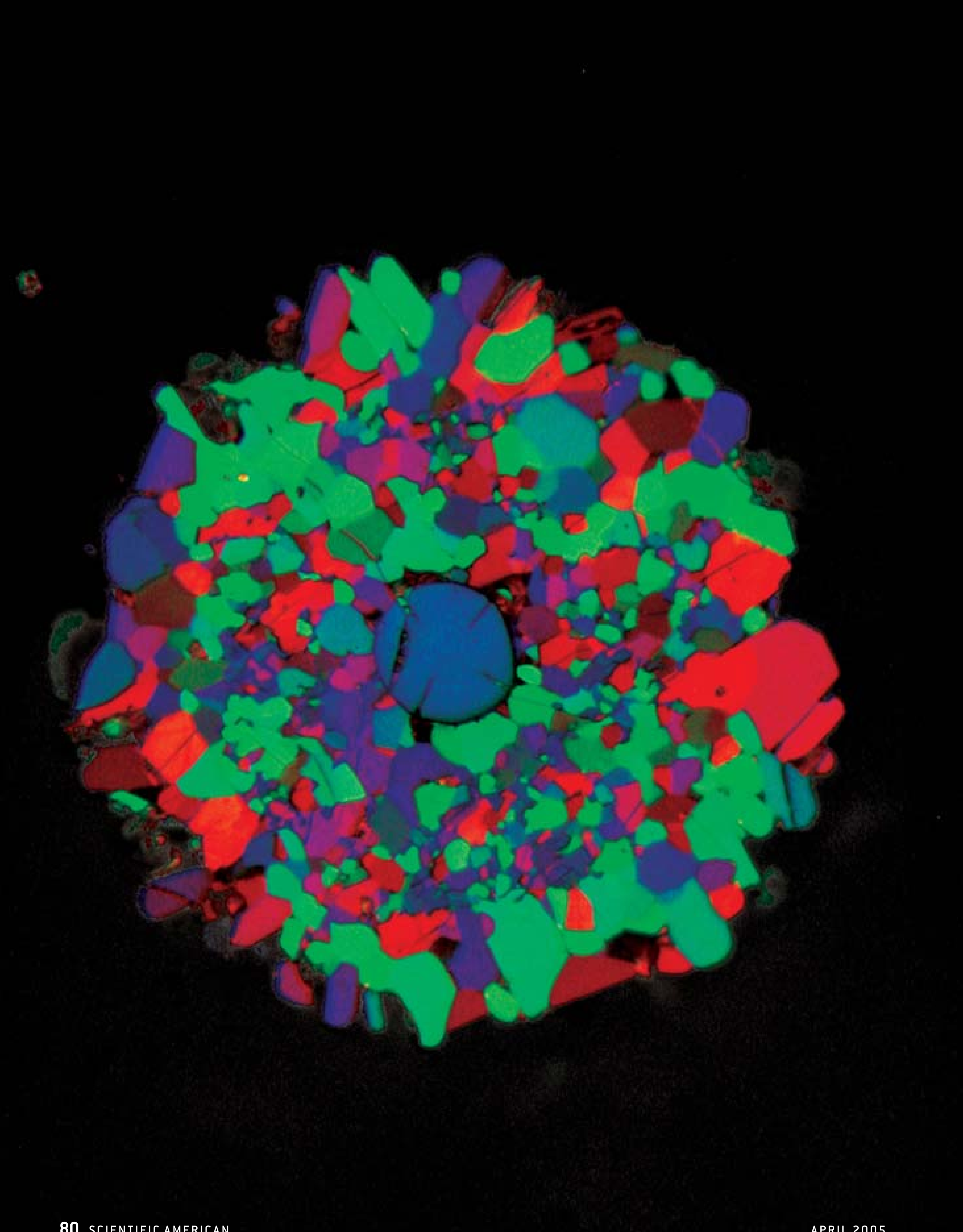
Infants as a Commodity in a Baboon Market. S. P. Henzi and L. Barrett in *Animal Behaviour,* Vol. 63, No. 5, pages 915–921; 2002.

Monkeys Reject Unequal Pay. Sarah F. Brosnan and Frans B. M. de Waal in *Nature,* Vol. 425, pages 297–299; September 18, 2003.

Living Links Center site: **www.emory.edu/LIVING_LINKS/**

Classic cooperation experiment with chimpanzees: **www.emory.edu/LIVING_LINKS/crawfordvideo.html**

JIM SUGAR *Corbis*



LOW-TEMPERATURE SUPERCONDUCTIVITY IS WARMING UP

Magnesium diboride defies the once conventional wisdom about what makes a good superconductor. It becomes superconducting near the relatively warm temperature of 40 kelvins—which promises a variety of applications

By PAUL C. CANFIELD AND SERGEY L. BUD'KO

HAL SAILSBURY *Materials Preparation Center, Ames Laboratory*

POLISHED CROSS SECTION of a magnesium diboride wire segment shows that the wire is essentially 100 percent dense and is made up of small, unoriented grains, which reflect light differently, giving rise to various colors. Such wires are useful for basic research into the material's superconductivity. The wire is 0.14 millimeter in diameter.

IMAGINE walking around in your backyard and suddenly discovering a vein of gold in a corner you thought you knew well. Or imagine how Jed Clampett of the *Beverly Hillbillies* felt when oil started bubbling up through the ground. A similar sensation of incredulous excitement swept over the solid-state physics community in the early weeks of 2001, when researchers announced that magnesium diboride (MgB_2) superconducts—conducts electricity without resistance—at temperatures approaching 40 kelvins.

This simple compound had been studied in the 1950s and had been on the shelves in some laboratories for various mundane purposes for decades, with no one suspecting its enormously valuable hidden talent. Although 40 K (or –233 degrees Celsius) may sound rather low, it was nearly double the record for compounds made of metals (about 23 K for niobium-based alloys, which are widely used in research and industry). A transition temperature that high can be achieved by technologies that cost much less than those needed to bring about superconductivity in the niobium alloys. Possible applications include superconducting magnets and power lines.

Unlike high-temperature superconductors (copper oxide materials that superconduct at temperatures as high as 130 K), MgB_2 seems to be a traditional superconductor, albeit a novel variant. In their decades-long quest for superconductors with ever higher transition temperatures, physicists had developed rules of thumb regarding what kind of combinations of elements to try. In addition, many suspected that 23 K was close to the maximum transition temperature possible for a traditional superconductor. To their great surprise, MgB_2 defied these rules and blew away the barrier to higher temperatures.

The speed with which understanding of MgB_2 grew was absolutely amazing. Jun Akimitsu of Aoyama Gakuin University in Tokyo announced the discovery at a meeting in mid-January 2001. Just two months later about 100 two-minute talks on the topic were presented at the American Physical Society's annual March meeting, and more than 70 research papers had been electronically posted on the arxiv.org preprint archive. This burst of activity happened for a few reasons. First, once you figure out how, it is fairly simple to make relatively pure MgB_2. Second, in 2001 the condensed-matter physics community was more wired together by the Internet than ever before. These two ingredients, combined with the promise of a new, simple superconductor with a high transition temperature, formed an explosive intellectual mixture.

Confirming the Discovery

AT FIRST, news of Akimitsu's announcement spread only by word of mouth and e-mail. No research paper or electronic draft was available. When the news reached our group a few days after the meeting, we asked a series of questions: Can we make high-purity, solid pieces of this stuff? (On the shelf, MgB_2 is a not so pure powder.) Does it really superconduct near 40 K? (There had been almost two decades' worth of USOs, or "unidentified superconducting objects"—compounds reported to have exceptionally high transition temperatures that other researchers could not replicate.) If MgB_2 does superconduct, can we uncover the mechanism of its superconductivity? And finally, can we delineate some of this compound's basic properties? Happily for one and all, the answer to each of these questions was yes.

The rumor of Akimitsu's discovery started a frantic and wonderful time for us and for other research groups. Our team specializes in studying the physical properties of metallic compounds, so as soon as we heard about the report, we emptied all of our furnaces of existing experiments and started trying to produce MgB_2.

Making the compound was a tricky business initially. It is an example of an intermetallic compound, one made of two or more metallic elements. The simplest way of making intermetallic compounds—by just melting the elements together—was not possible in this case, because the two elements have very different melting points: 650 degrees C for magnesium and higher than 2,000 degrees C for boron. Because magnesium boils at just over 1,100 degrees C, the magnesium would evaporate before the compound could form, or, in the vernacular, the magnesium would grow legs and walk away.

But the vaporization of magnesium suggested an alternative method: we could seal a piece of magnesium and some powdered boron inside a tantalum vessel, which is inert, and subject them to a temperature high enough to melt but not to boil the magnesium (say, 950 degrees C). Magnesium has a relatively high vapor pressure—indeed, one third of an atmosphere of magnesium vapor exists in equilibrium with the liquid metal at 950 degrees C. We expected that this dense vapor would diffuse into the solid boron, producing pellets of MgB_2. Sure enough, we found that in as little as two hours this process produced very high purity MgB_2 in the form of a loosely sintered pellet (like sandstone). Within three days of hearing the rumors, we had made these pellets and were able to confirm superconductivity at near 40 K.

Having figured out how to make MgB_2 and confirmed that it is a superconductor, we asked the next burning question: Was it an old-fashioned superconductor whose behavior could be explained by a long-established theory called BCS theory (from the initials of its three discoverers' last names) or an example of a more exotic type [*see box on page 86*]? If it was an exotic type, that would be a profound scientific discovery. On the other hand, if it was a conventional BCS superconductor, the exceptionally high transition temperature would demand an explanation, but the prospects for using the material in applications would be more encouraging.

For several reasons, some researchers thought that MgB_2 was not a standard BCS superconductor. First, before high-temperature superconductors were discovered in 1986, two decades had gone by with the highest transition temperature stuck at around 20 K. This fact led some theorists to suggest that about 30 K was the maximum temperature possible for superconductivity in compounds that obey BCS rules. The high-temperature copper oxide superconductors far exceeded that limit, but they are not thought to be BCS superconductors.

Overview/*Magnesium Diboride*

- In 2001 researchers discovered that the seemingly unexceptional compound magnesium diboride superconducts below about 40 kelvins, nearly twice the temperature of other similar superconductors. Its maximum practical working temperature is about 20 to 30 kelvins.
- That temperature can be achieved by cooling with liquid neon or hydrogen or by closed-cycle refrigeration, all of which are much less expensive and troublesome than the liquid-helium cooling required by the niobium alloys that are widely used in industry at about four kelvins.
- When doped with carbon or other impurities, magnesium diboride roughly equals or betters niobium alloys at retaining its superconductivity in the presence of magnetic fields and when carrying an electric current. Potential applications include superconducting magnets, power lines and sensitive magnetic field detectors.

Second, MgB_2's relatively high transition temperature, or critical temperature (T_c), violated one of the old rules of thumb in the search for intermetallic compounds having a higher T_c: the more electrons that could participate in the phase transition to the superconducting state, the higher the transition temperature would be. Neither magnesium nor boron brought particularly many electrons to MgB_2.

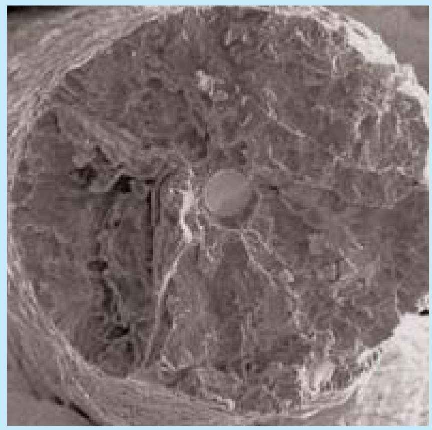
A very direct experimental test can tell whether a superconductor is following the BCS theory. A key role in the theory is played by lattice vibrations. Imagine that the heavy positive ions of the crystal lattice are held in place by strong springs (the chemical bonds). Excitations such as heat manifest as vibrations of the ions at characteristic frequencies. BCS theory predicts that the transition temperature of a superconductor is proportional to the frequency of its lattice vibrations. As is the case with everyday objects such as wineglasses or guitar strings, objects made from lower-mass materials have higher characteristic frequencies than otherwise identical objects made from higher-mass materials. By using a different isotope of magnesium or boron, we can make MgB_2 out of atoms of different mass, which will alter the lattice vibration frequency, which in turn should alter T_c in a specific way.

Boron has two stable, naturally occurring isotopes: boron 10 and boron 11. The simplest prediction of the BCS model is that T_c should differ by 0.85 K for two samples of MgB_2 made with pure boron 10 and boron 11. With our first sintered pellets of MgB_2, we discovered a shift of 1 K. The fact that the shift in T_c was a little larger than the simple prediction can be accommodated by BCS theory—it indicates that the boron vibrations are more important to the superconductivity than the magnesium vibrations [*see box on page 87*].

The closeness of this shift to the predicted 0.85 K revealed that MgB_2 is most likely a BCS superconductor, albeit an extreme example that has a much higher transition temperature than any other. The predictions of an approximate 30 K upper limit to BCS superconductivity were apparently not valid. This was good news, because standard intermetallic BCS superconductors are much easier to work with and can form useful wires much more readily than copper oxide–based superconductors can. Indeed, it suddenly dawned on our group that we could form MgB_2 wires by simply exposing boron filaments to magnesium vapor [*see box above*]. Such wires are of greater use than sintered pellets for many measurements and for applications such as magnets.

Making Wires

WIRES were made by reacting magnesium vapor and boron filaments.



CROSS SECTION of a snapped magnesium diboride wire segment reveals a central core of tungsten boride, 0.015 millimeter in diameter.

NORMAN E. ANDERSON, JR. (*top*), AND HAL SAILSBURY *Materials Preparation Center, Ames Laboratory* (*bottom*)

Within a couple of weeks after the announcement of superconductivity in MgB_2, we had devised a technique for making wire segments of this remarkable superconductor. MgB_2 can be formed by reacting magnesium vapor with boron, a process that can take place within hours at temperatures near 1,000 degrees Celsius: the boron essentially sucks the magnesium vapor out of the environment and becomes MgB_2 (swelling up dramatically in the process). Imagine a dry sponge sucking water vapor out of the air on a humid day. This process works with boron fibers that can be purchased in lengths of hundreds of meters; it has been applied to filaments with starting diameters ranging from 0.1 to 0.3 millimeter.

Wire segments such as these are very useful for basic research, allowing for the measurement of the intrinsic physical properties of MgB_2. Before such wire segments could be used in practical applications, they would need to have a conductive, malleable sheath to provide structural support. (The conductive sheath also carries the current should the superconductivity fail, preventing catastrophic heating of the MgB_2.) A suitable sheath has not yet been developed.

A more common method of wire synthesis is called "powder in a tube." This technique involves pouring powdered magnesium and boron or powdered MgB_2 into a tube, drawing the tube into a wire, and then reacting or annealing the wire to produce a solid structure. This technique has produced research samples that range from tens to hundreds of meters long.

Despite MgB_2 being a rather young superconductor, companies have been taking notice and are working toward commercializing it. Examples include Diboride Conductors and Hyper Tech Research, which are small companies focusing on making and improving the properties of MgB_2 wires, and Specialty Materials, which is a larger materials company with expertise in producing boron filaments.

—P.C.C. and S.L.B.

A History of Superconductivity

Heike Kamerlingh Onnes discovered superconductivity in 1911 when he used liquid helium as a coolant to study the electrical properties of metals at low temperatures. To everyone's surprise, when mercury was cooled to about 4.2 kelvins, it suddenly lost all electrical resistance. This threshold is known as the critical temperature, or T_c.

Other materials having ever higher critical temperatures were discovered slowly but surely during the first five decades of superconductivity research. All these superconductors were either pure metallic elements or intermetallic compounds (made of two or more metallic elements). But from the 1960s through the mid-1980s the maximum value of T_c seemed to be stuck in the low 20s.

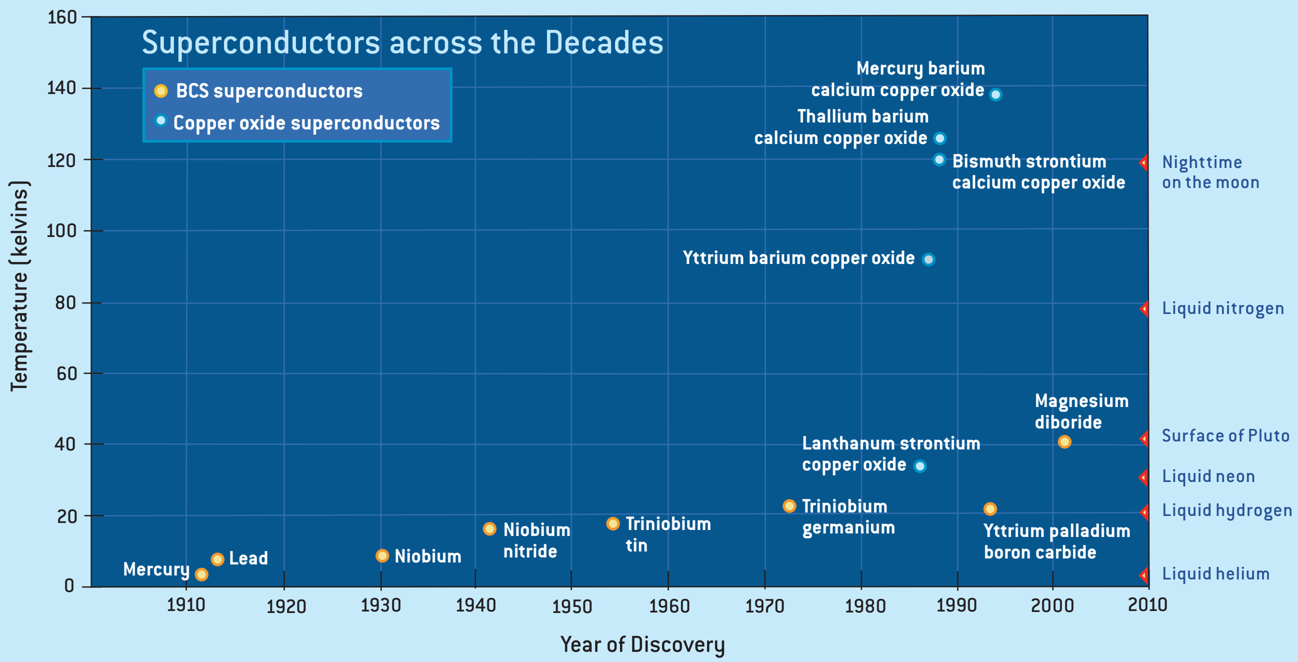
All this changed in 1986 with the discovery of high-temperature superconductivity in a slew of copper oxide–based compounds. During the first few years after this discovery, T_c values shot up, with mercury-barium-calcium-copper-oxide having a T_c of about 130 K. This was a fantastically exciting time, but it soon became clear that the leading theory of how superconductivity arises—known as BCS theory [*see box on page 86*]—does not explain the absence of resistance in these materials. Despite almost 20 years' worth of effort, there is still no definitive theory of how or why the copper oxide compounds superconduct.

These compounds also pose a multitude of physical challenges. Initially they were hard to make in either high-purity or single-crystal form, making the measurements of their fundamental properties difficult. In addition, the synthesis of wires is not easy: unlike the intermetallic superconductors, the individual grains that make up a piece of one of these oxides have to be aligned with respect to one another for the wire to have useful engineering properties. These problems left researchers and engineers wishing for a substance with the somewhat easier material properties of the intermetallic superconductors that also had a critical temperature significantly higher than 20 K.

By the dawn of the new millennium, then, the superconducting state could be achieved with varying degrees of ease and expense. In the oxides, superconductivity was practical near 77 K, which can be reached relatively easily by bathing the material in liquid nitrogen. The older intermetallic compounds such as triniobium tin were being used in the laboratory and as medical magnets operating at temperatures closer to 4 K, which can be reached with liquid helium.

The discovery in 2001 that the simple intermetallic compound magnesium diboride superconducts at 40 K, about double the temperature of the other intermetallics, was almost exactly what the doctor (or in this case, engineers) had ordered.

—P.C.C. and S.L.B.



TOMMY MOORMAN AND CARLI MORGENSTEIN

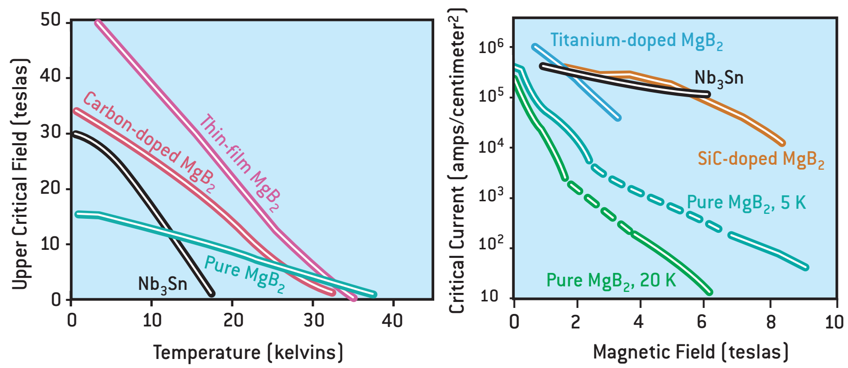
Uses of Superconductors

ALTHOUGH IT OCCURS only at very low temperatures, superconductivity has a wide variety of present-day uses, as well as potential future applications. Some of the most obvious derive from superconductors' ability to carry high currents with no energy losses or resistive heating. An example is superconducting magnets that can produce magnetic fields in excess of 20 tesla (about 500 times stronger than a typical refrigerator magnet). Low-temperature superconducting magnets such as these (and less strong ones) are used in labs and in magnetic resonance imaging machines in hospitals. Sales of these magnets, crafted from niobium-based compounds and alloys, continue to grow.

Another high-current application that has been proposed is lossless power transmission lines, which can carry

IMPROVING PERFORMANCE

Maintaining superconductivity in a magnetic field and when carrying a current is crucial for applications. The plotted data show how doping with impurities has improved the performance of MgB_2; it now equals or exceeds that of the industrially favored triniobium tin (Nb_3Sn). The graph at the left shows that wire segments of carbon-doped MgB_2 and a thin film of MgB_2 with an unknown level of impurities withstand a higher magnetic field ("upper critical field") than Nb_3Sn at all temperatures. The data at the right (taken at about 4 K except where noted) show that MgB_2 doped with silicon carbide (SiC) equals the current-carrying capacity of Nb_3Sn, but other variants are significantly inferior. Dashed lines are interpolations.



LUCY READING-IKKANDA

much higher current densities than nonsuperconducting ones. To date, researchers have successfully tested several copper oxide–based prototypes that have been cooled to near 70 K with liquid nitrogen.

Generally speaking, to act as superconductors in practical applications, compounds need to be cooled substantially below their T_c, to about 0.5 to 0.7 T_c, because large electric currents or strong magnetic fields destroy the superconductivity closer to T_c. Consequently, a T_c of 20 K may imply an operating temperature of 10 K, which means that the superconductor has to be chilled by liquid helium, a costly and somewhat difficult option.

The applied research community is interested in MgB_2 because this material can be cooled to viable operating temperatures more easily than the lower-T_c niobium-based alloys and compounds that are employed today. MgB_2 can be cooled by liquid hydrogen or liquid neon or by fairly cheap, closed-cycle refrigerators that can readily reach below 20 K.

But if this vision is to become a reality, MgB_2 will need to have good superconducting properties. Researchers are paying particular attention to the superconductor's mixed phase, in which a magnetic field partially destroys the superconductivity—in most real applications, the material will be in this phase. Weak magnetic fields do not produce the mixed state; the superconductor excludes such fields from its interior and remains superconducting. At intermediate fields, however, the material allows the magnetic field to penetrate in the form of small tubes of magnetic flux known as vortices. The insides of these tubes are nonsuperconducting, but outside of them the material remains superconducting.

This mixed phase still manifests many of superconductivity's useful characteristics. As the strength of the applied magnetic field increases, the percentage of the material occupied by the flux tubes increases until they overlap fully, at which point the whole material is nonsuperconducting. The field strength at which superconductivity is lost is referred to as the upper critical field and is a key property that determines how useful a superconductor will be in practice.

Most applications will involve intermediate fields (the field is strong enough to be useful but not so strong as to destroy superconductivity altogether), so the goal becomes maximizing the range of temperatures and magnetic fields in which the superconducting mixed phase survives. Temperature also plays a role in these considerations because the upper critical field of a superconductor varies with temperature. Just below T_c, the upper critical field is close to zero—that is, even the weakest field destroys the superconductivity. At lower temperatures the superconductivity can resist stronger fields [*see box above*].

Fortunately, the upper critical field of a material can be tuned by making the compound in differing ways, generally by adding certain impurities. For example, when some carbon is substituted for boron in MgB_2, the upper critical field is dramatically improved. Our group and others have shown that for about a 5 percent substitution of carbon, the upper critical field of MgB_2 can be more than doubled—a fantastic and important improvement in bulk samples.

In addition, the group of David C. Larbalestier at the University of Wisconsin–Madison has shown that thin films of MgB_2 have even higher values of the upper critical field, well above those of triniobium tin (Nb_3Sn). The thin-film data present a vital mystery: What is giving rise to the high values? Is it small amounts of oxygen? Is it some other element sneaking

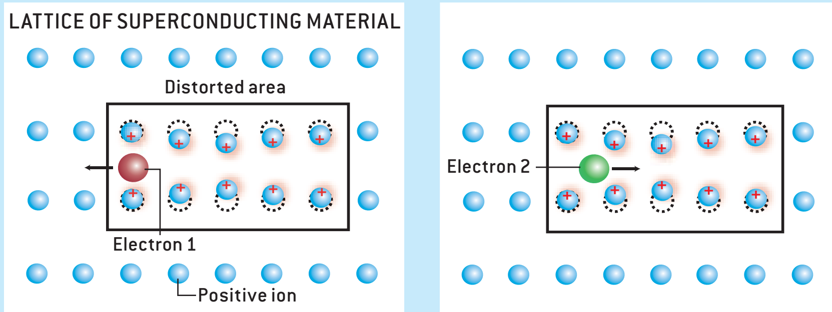
THE AUTHORS

PAUL C. CANFIELD and *SERGEY L. BUD'KO* both work at the Department of Energy's Ames Laboratory in Iowa. Canfield is also a professor of physics and astronomy at Iowa State University. His research focuses on the design, discovery, growth and characterization of novel materials and phenomena, primarily focusing on the low-temperature electronic and magnetic states of metallic compounds. Bud'ko's research interests include thermodynamic, magnetic and transport properties of novel materials; quantum oscillations in metals and semimetals; and the physical properties of materials in extreme conditions that combine high pressure, strong magnetic field and low temperature. The authors gratefully acknowledge their fruitful collaborations with R. Wilke, D. Finnemore, C. Petrovic, G. Lapertot, M. Angst, R. Ribeiro and N. Anderson. Their work was supported by the Director for Energy Research, Office of Basic Energy Sciences.

Predictions of BCS Theory

In 1957 physicists John Bardeen, Leon N. Cooper and J. Robert Schrieffer proposed an explanation of the mechanism underlying superconductivity of metals in a theory that bears their initials. In a normal, nonsuperconducting metal, electrons scatter off of defects and imperfections, which generates resistance. According to BCS theory, superconductivity takes place when the electrons instead act as a single extended collective object that can move without scattering.

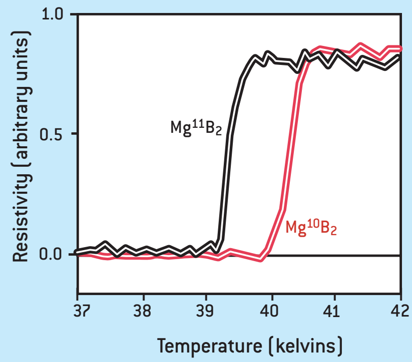
The building blocks of this new electronic state are pairs of electrons, called Cooper pairs, in which the partners are weakly attracted to each other. This attraction between two like-charged particles, at first sight seemingly impossible, occurs because the metal is composed of positively charged ions as well as electrons. As one member of the Cooper pair moves through the metal, it leaves a positively charged ionic distortion in its wake. This fleeting net positive charge attracts a second electron. In this way, the lattice distortion loosely couples the electrons. (More precisely, lattice vibrations of a specific frequency are involved in the coupling.) A rather gross analogy is that of two kids bouncing on a large trampoline. Even though there is no direct attraction between the kids, they will tend to bounce toward each other because of the distortion that the other causes in the tarp beneath their feet.



FORMATION OF ELECTRON PAIRS known as Cooper pairs (*above*) ultimately leads to superconductivity. One electron leaves in its wake a distortion of the lattice of positively charged ions in a metal (*left panel*). Shortly thereafter, the second electron is attracted by the resulting concentration of positive charge (*right panel*). In effect, the two electrons are weakly attracted to each other.

The Cooper pairs overlap one another, and below the critical temperature (T_c) they form an extended electronic state that no longer experiences any electrical resistance.

A simplified version of the BCS theory predicts that T_c depends on three properties of the material in question: the number of electrons that can participate in the superconducting state (the more electrons that can participate, the higher the T_c); the characteristic frequency of lattice vibrations involved in coupling the electrons in the Cooper pair (the higher the frequency, the higher the T_c); and the strength of the coupling between the lattice distortion and the electrons (the stronger this coupling, the higher the T_c). For decades, the search for higher T_c values focused on optimizing these three related properties, with a preference for trying to improve the first two. MgB_2 seems to have a high T_c because of its stronger electron-lattice coupling, the third property. —*P.C.C. and S.L.B.*



ELECTRICAL RESISTIVITY of MgB_2 drops to zero as the material is cooled below its critical temperature of about 40 K. The critical temperature is different for samples made with pure boron 10 and boron 11. This clear isotope shift is predicted by BCS theory and thus indicates that the superconductivity in MgB_2 is traditional BCS superconductivity.

LUCY READING-IKKANDA

in and doping in unknown ways? Is it strain in the structure of the MgB_2 in the films? Whatever the answers to those questions, clearly MgB_2 is a promising material for superconducting magnets that can function at higher temperatures and possibly even in higher fields than triniobium tin, which is currently the preferred compound for such magnets.

The second superconducting property of particular interest for applied physics is the critical current density. This quantity delineates the maximum amount of current that a superconductor can carry and still maintain zero resistance. For current densities above the critical current density, the vortices (the small nonsuperconducting regions of the sample) start to slip or move. Once these regions start moving, energy losses occur—that is, the material has a nonzero resistance. To counter this effect, the vortices can be pinned (in essence, nailed down) by introducing the right type of defect into the superconductor. Often the vortex pinning can be increased by making the individual crystallites (or grains) of the material smaller, thus increasing the surface area associated with grain boundaries, where vortices get pinned. Another method of increasing vortex pinning involves adding microscopic inclusions of some second material such as yttrium oxide or titanium diboride.

Currently one of the major challenges associated with making MgB_2 a useful superconducting material is to increase its critical current density at higher magnetic fields. The critical current density of pure MgB_2 is comparable to that of triniobium tin at low magnetic fields but falls off much more rapidly at higher fields. This is not good news if the goal is to use MgB_2 in magnets, which are meant to produce a strong field. On the other hand, in the four years since the discovery of superconductivity in this compound, the research community has made considerable improvements in critical current density, both in the low-field value and, perhaps more important, in the higher-field values. Research in this area is very active, and it appears that physicists will soon make further improvements and achieve a better understanding of what will provide a good pinning site in MgB_2.

Past, Present and Future

THE DISCOVERY of superconductivity in MgB_2 is simultaneously the fruition of decades of focused research and a stark reminder that nature does not always heed the rules of thumb we make up in our often vain attempts to describe her. Although MgB_2 was known to exist for about 50 years, it was never tested for superconductivity, partly because it did not fit our image of a likely intermetallic superconductor. Luckily, in the search for new materials and properties, nature's voice can still be heard over the din of our prejudices.

Over the past four years, humankind's understanding of superconductivity in MgB_2 has evolved at breakneck speed. We have a clear idea of the properties of high-purity MgB_2, and we are learning how to modify the material so as to improve the ranges of magnetic field and current density over which it can be useful. The properties at 20 to 30 K have improved to the point that it appears high-current-density applications, such as magnets, can be made to operate either with cryogens such as liquid hydrogen or liquid neon or with closed-cycle refrigerators. Prototype coated wires and even some initial magnets have been made, but more work is needed to optimize the superconductor's properties and to understand its metallurgy as well as that of possible wire-coating materials.

On the whole, the future for MgB_2 looks quite promising. Indeed, if a shift toward a hydrogen-based economy occurs, then MgB_2 could truly come into its own. If large quantities of hydrogen are to be produced, for example, at small pebble-bed reactors [see "Next-Generation Nuclear Power," by James A. Lake, Ralph G. Bennett and John F. Kotek; SCIENTIFIC AMERICAN, January 2002], the hydrogen will have to be transported in some manner. One way would be through insulated, liquid-carrying pipes that would maintain temperatures below hydrogen's 20 K boiling point. These pipes could constitute the cryogenic system for lossless power cables made of MgB_2 sharing the space inside the thermal insulation. Although such a system currently sounds more like science fiction than an engineering reality, it has been proposed for serious study.

After the discovery of the first copper oxide–based superconductor, researchers found scores of other superconducting copper oxides. Yet four years after the discovery of MgB_2, no other related compounds have been found to have anomalously high T_c values. The discovery of superconductivity in the oxides was akin to discovering a whole new continent (with wide expanses to be explored). The discovery of superconductivity in MgB_2, on the other hand, was more like the discovery of an outlying island in a well-explored archipelago. We do not know if this is the final member of the chain or if yet another surprise awaits us out there. SA

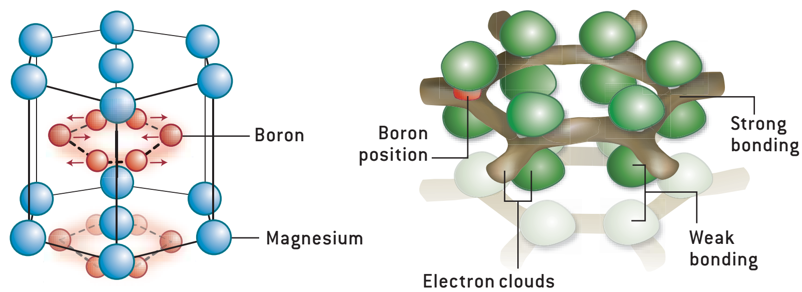
Structure and Bonding

One of the primary reasons for the surprisingly high transition temperature of MgB_2 is the strength of the interaction between certain electrons and certain lattice vibrations. The strong interaction arises because of the material's structure and bonding.

The boron atoms in MgB_2 form a hexagonal honeycomb pattern (*red, left*). In MgB_2 these layers are separated by layers of magnesium (*blue*). The electrons responsible for the ordinary electrical conductivity and also the superconductivity are associated with the boron layers and are involved in one of two different types of bonding in the material (*right*). A very strong bonding occurs within the hexagonal plane, and a much weaker bonding occurs between the boron layers.

The conduction electrons of the planar bonds are very strongly affected by in-plane lattice vibrations (*arrows, left*). This strong interaction, or coupling, results in a state that remains superconducting at higher temperatures.

MgB_2 has revived a very exciting basic physics question: Can a superconductor have superconductivity involving two distinct collections of electrons (*green* and *gold*) forming two distinct seas of Cooper pairs? Experimental evidence suggests that this is the case in MgB_2, which would be the first clear example of this phenomenon. *—P.C.C. and S.L.B.*



LUCY READING-IKKANDA

MORE TO EXPLORE

Superconductivity at 39 K in Magnesium Diboride. Jun Nagamatsu et al. in *Nature,* Vol. 410, pages 63–64; March 1, 2001.

Magnesium Diboride: One Year On. Paul C. Canfield and Sergey L. Bud'ko in *Physics World,* Vol. 15, No. 1, pages 29–34; January 2002.

Energy for the City of the Future. Paul M. Grant in *Industrial Physicist,* Vol. 8, No. 1, pages 22–25; February/March 2002. Available at **www.aip.org/tip/INPHFA/vol-8/iss-1/p22.pdf**

Magnesium Diboride: Better Late than Never. Paul C. Canfield and George W. Crabtree in *Physics Today,* Vol. 56, No. 3, pages 34–40; March 2003.

Superconductivity in MgB_2: Electrons, Phonons and Vortices. Edited by Wai Kwok, George W. Crabtree, Sergey L. Bud'ko and Paul C. Canfield. *Physica C,* Vol. 385, Nos. 1–2; March 2003.

A TOXIN AGAINST PAIN

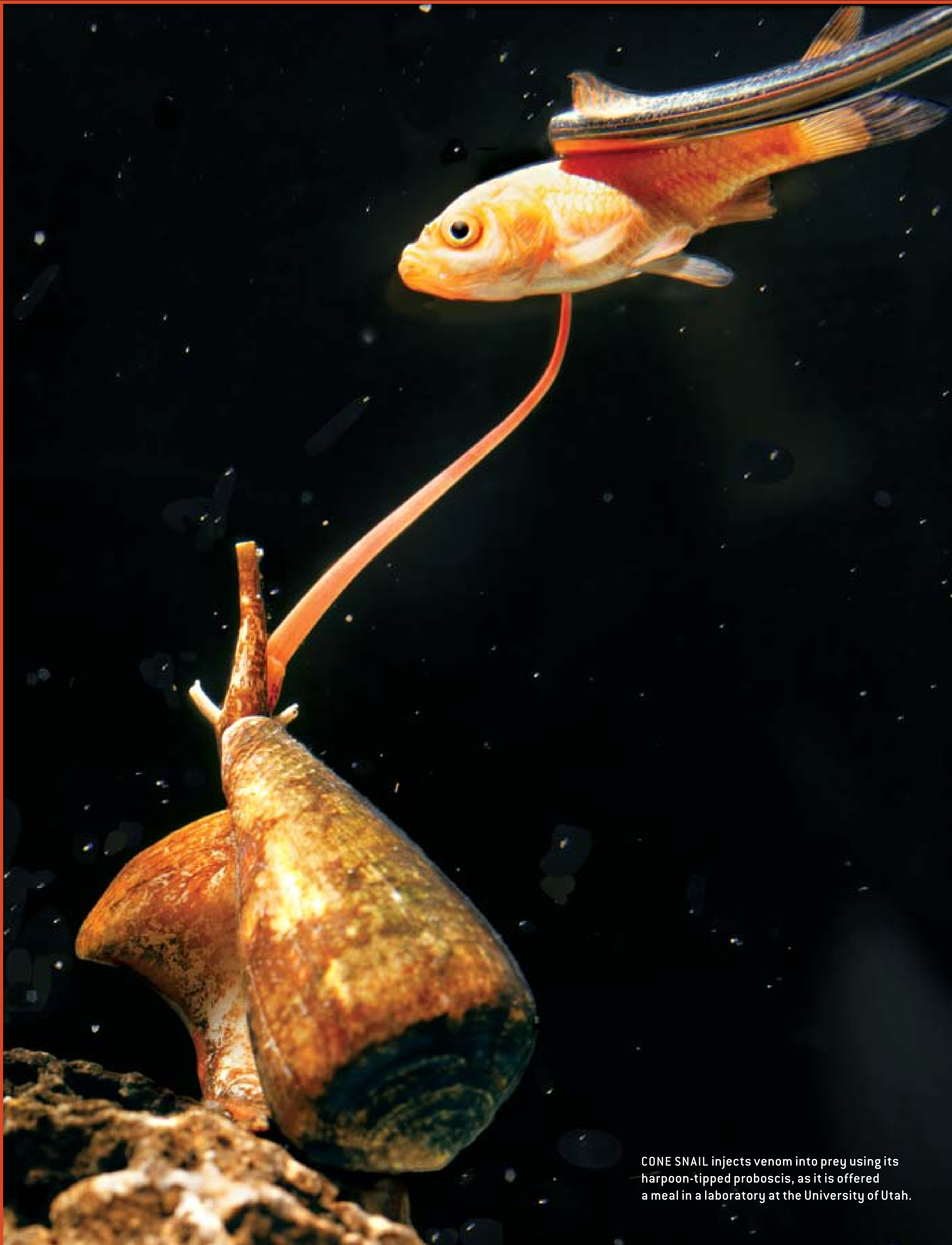
By Gary Stix

For years, scientists have promised a new wave of drugs derived from sea life. A recently approved analgesic that is a synthetic version of a snail toxin has become one of the first marine pharmaceuticals

The past 18 months were very good ones for hypnotists, yoga teachers and acupuncturists. For many chronic pain sufferers, promises of relief from various forms of alternative medicine seemed like rational options amid the unending stream of negative reports about Vioxx, Celebrex, Aleve and Rush Limbaugh's addiction to painkillers.

Not all was lost for patients who prefer medicine to meditation. With little fanfare, the Food and Drug Administration approved in late December two new drugs intended to treat a form of pain that often proves resistant to anti-inflammatories and opiates—the two predominant classes of pharmaceuticals for analgesia. Medical specialists welcomed their arrival. "It's an embarrassment that we're treating pain with opiates and aspirinlike compounds," notes Edwin McCleskey of the Oregon Health and Sciences University. "Opiates are more than 2,000 years old, and aspirin is nearly 200 years old."

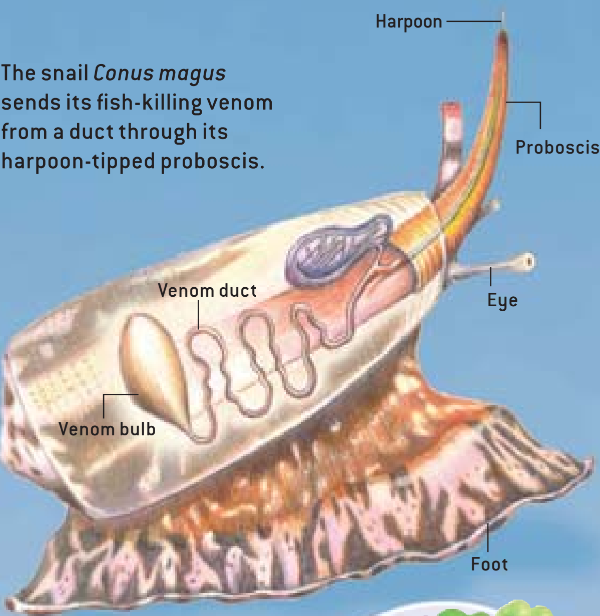
BRUNO GAGET



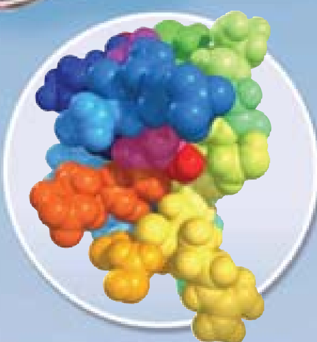
CONE SNAIL injects venom into prey using its harpoon-tipped proboscis, as it is offered a meal in a laboratory at the University of Utah.

FROM SNAIL TO PATIENT

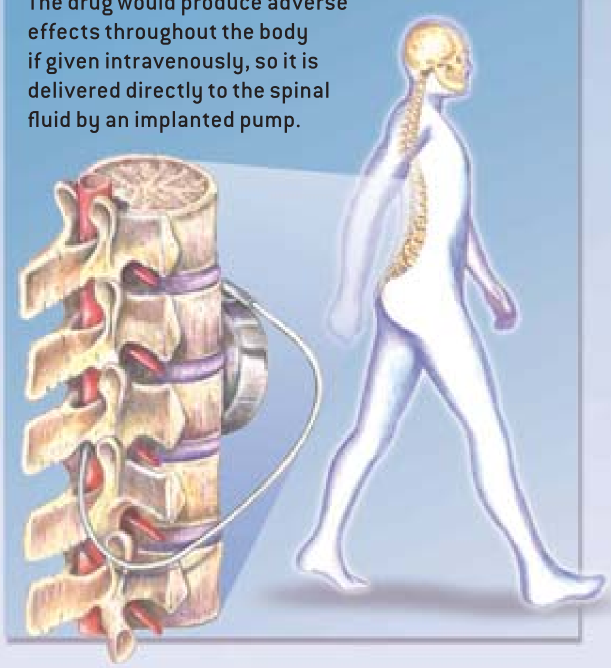
The snail *Conus magus* sends its fish-killing venom from a duct through its harpoon-tipped proboscis.



That venom turned out to include a substance (*right*), known as an omega-conopeptide, that can alleviate pain. Researchers now synthesize the conopeptide to make the drug Prialt, generically known as ziconotide.



The drug would produce adverse effects throughout the body if given intravenously, so it is delivered directly to the spinal fluid by an implanted pump.



Intractable pain often gets addressed with drugs first approved for other conditions. Pfizer's Lyrica (pregabalin) received FDA endorsement on December 31 for neuropathic pain caused by nerve damage resulting from diabetes and shingles. Anticonvulsants, a class that includes Lyrica, have been used for palliation, sometimes without regulatory approval. Last year Pfizer agreed to pay $430 million in criminal and civil penalties because its Warner-Lambert division had illegally promoted an epilepsy drug, Neurontin, for neuropathic pain and other unapproved uses before it was acquired by Pfizer in 2000.

The other such analgesic that the FDA approved in December has none of Lyrica's blockbuster potential. Prialt (ziconotide) from Elan requires that a pump be implanted or used externally to send the drug by catheter into the spinal fluid, a technology often reserved to deliver morphine to critically ill AIDS and cancer patients. Prialt may not dazzle Wall Street, but from the perspective of neuroscientists and pharmacologists, it is by far the more interesting of the two compounds. "It could be argued that this is the first drug for pain that went from animal tests into patients, rather than going from a patient using it for something else back into animals to get validated for pain," remarks Allan I. Basbaum, a professor of anatomy at the University of California at San Francisco. "It's a proof of principle that there really are new drug targets worth going after."

Prialt, a synthetic copy of a toxin from the Magician's cone snail, *Conus magus,* a mollusk from the Indo-Pacific region, is also one of the first pharmaceuticals that demonstrates the promise that marine life, particularly invertebrates, holds for drug developers.

The route to Prialt began in the early 1970s, when Baldomero Olivera, who had recently completed postdoctoral work at Stanford University, returned to his native Philippines to set up a laboratory. At Stanford, Olivera had helped isolate and purify DNA ligase, the enzyme that joins pieces of DNA. He wanted to continue his research on the enzyme in the Philippines but was unable to procure the necessary equipment. A shell collector, he wondered whether poisonous cone snails might contain molecules that could block nerve channels—and that could be used by neuroscientists in the same way the toxins from puffer fish or a Taiwanese snake are. "I got started without any grand vision," Olivera recalls. "I was mainly looking for a project to work on productively."

Eventually he made his way to a teaching position at the University of Utah and planned to sideline studies on snail toxins in favor of his previous work on DNA. In 1978, a few years after Olivera's return, Craig Clark, a 19-year-old undergraduate working in his lab, showed an interest in the snail toxins. Clark wanted to test what would happen if just one or a few of the 100 or so peptides that make up the venom of a highly lethal cone snail, *C. geographus,* were injected directly into the brains of mice instead of into their abdomens, the protocol for earlier experiments. Olivera was not optimistic but let Clark proceed. To everyone's astonishment, the pep-

LUCY READING-IKKANDA; KEN EWARD (*inset, middle*) tides elicited a variety of behaviors. One peptide made a mouse sleep; another made it shake. Still another induced scratching.

A growing recognition of the diversity of cone snails and cone snail toxins—at least 50,000 peptides produced by 500 species (compare that with the 10,000 alkaloids identified in all plants)—prompted Olivera to abandon the DNA work. He then devoted himself to pursuing where this evolutionary variation came from in a comparatively short 50 million years. He was also trying to understand how ion channels function in the nervous system and how these toxins might be deployed in neuroscientific studies and drug development. "The snails are nature's neuropharmacists," he observes.

In Search of Venom

THE PUBLICATIONS FROM Olivera's group intrigued George Miljanich, a biochemist at the University of Southern California who was studying the transmission of neural signals across synapses, the contact points between neurons. Miljanich was involved in helping to identify and classify different types of calcium channels that convey chemical signals to cells in the nervous system. He had received a grant from the National Institutes of Health to try to develop cone snail toxins as probes to determine the function of different molecular pathways. His inability to get more than a drop of the precious venom, procured only after painstakingly milking a snail in Olivera's lab, made a job offer in 1988 from a start-up biotechnology company especially enticing. The company, Neurex, which was formed in 1986 by two Stanford professors with the goal of melding biotechnology and neuroscience, had also hired top-notch peptide chemists from U.C.S.F., where Miljanich had worked while a postdoctoral student.

Miljanich persuaded colleagues in Neurex's research department to undertake the difficult task of synthesizing omega-conopeptides—toxins that block certain calcium channels. The channels normally respond to a change in voltage across a cell membrane by allowing an influx of calcium ions, thereby facilitating transmission of a chemical signal across the synapse between nerve cells. Some omega-conopeptides come from *C. magus,* which is toxic to fish but not to humans. Originally skeptical about Miljanich's project, the Neurex management lost its reticence when its earlier, unfocused goal of isolating medicinally useful peptides from cow brains hit a dead end. Within a short time, Neurex was able to manufacture synthetic omega-conopeptides in gram quantities.

By then, Miljanich had made a list of possible uses of drugs that might be developed from one especially interesting conopeptide, first isolated by Olivera's laboratory. Analgesia was not first on the list but moved up fast. Though promising in test-tube studies as a possible epilepsy treatment, the favored omega-conotoxin, dubbed SNX-111 (and later Prialt/ziconotide), proved an utter flop when tested in mice. It actually induced shaking. The next item on the list called for administration of the drug intravenously to protect brain cells against the damage that occurs from lack of oxygen during stroke or head trauma. The company began the first phase of a neuroprotection clinical trial in 1993, but the FDA shut down the study temporarily when SNX-111 caused blood pressure to drop.

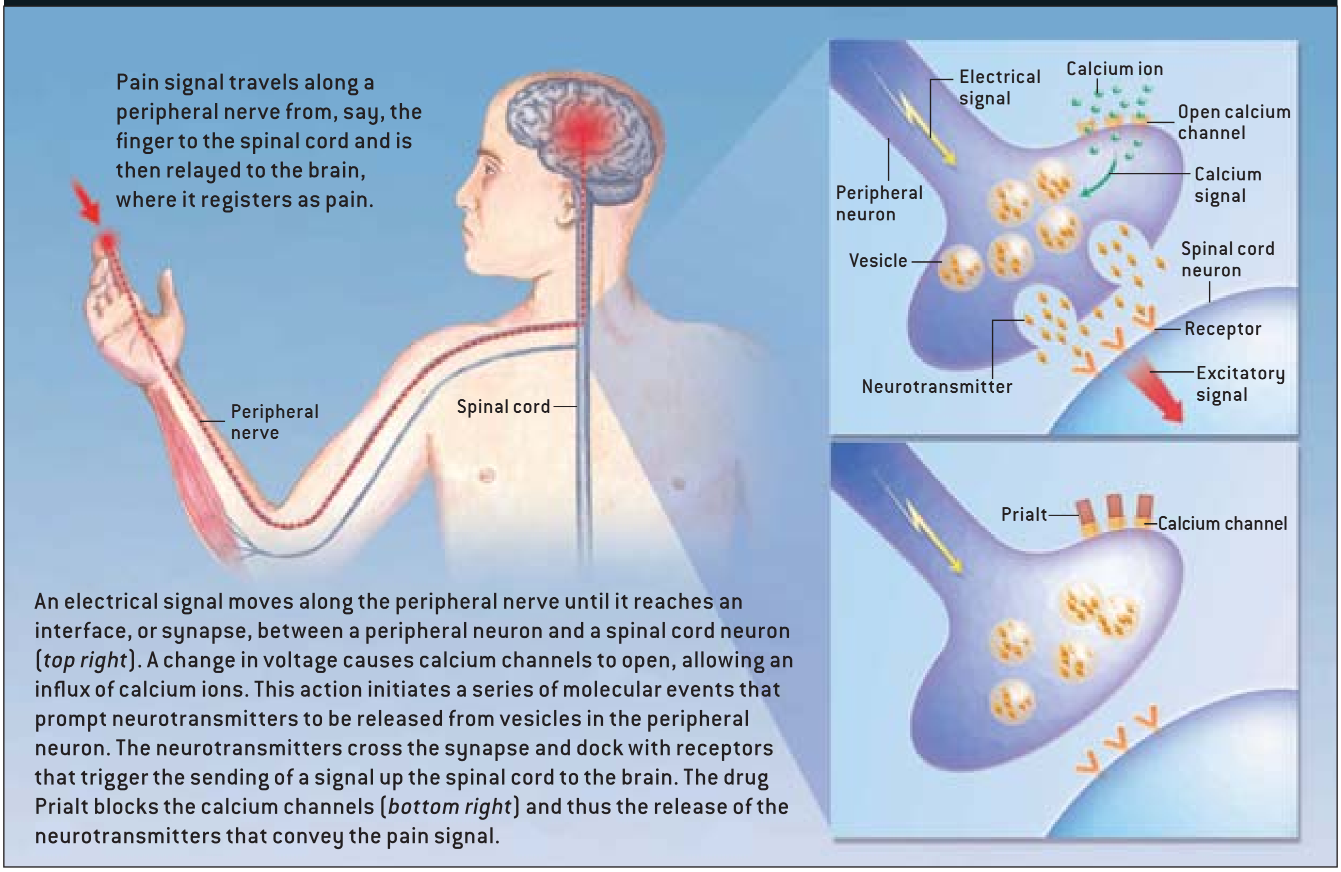
Getting desperate as it burned through cash, Neurex started preparing a clinical trial of SNX-111 for severe pain. The company's scientists had postulated that the compound might

MAKING DRUGS FROM CONE SNAIL TOXINS

COMPANY	DRUG	STAGE OF CLINICAL TRIAL	MOLECULAR DRUG TARGET	CONE SNAIL SPECIES
Cognetix Salt Lake City, Utah	CGX-1160, for chronic pain	Phase I, for cancer; Phase II, for spinal cord injury	Neurotensin receptor	*Conus geographus* (fish eater)
Elan Dublin, Ireland	Prialt (ziconotide)	Approved in the U.S. and Europe	Voltage-sensitive calcium channel	*C. magus* (fish eater)
Metabolic Pharmaceuticals Melbourne, Australia	ACV1, for neuropathic pain and for accelerating recovery of injured neurons	Has yet to enter human trials	Nicotinic acetylcholine receptor	*C. victoriae* (mollusk eater)
Xenome Brisbane, Australia	Xen 2174, for diminishing chronic pain	Phase I	Norepinephrine transporter (removes the neurotransmitter norepinephrine from synapses)	*C. marmoreus* (mollusk eater)
	TIA, for benign prostatic hyperplasia and for neurodegenerative and cardiovascular disorders	Has yet to enter human trials	Alpha-1-adrenergic receptor	*C. tulipa* (fish eater)

Phase I clinical trials are designed to assess the safety and tolerability of a drug and its effects in the body. Phase II trials examine drug effectiveness and safety.

TARGETED RELIEF



An electrical signal moves along the peripheral nerve until it reaches an interface, or synapse, between a peripheral neuron and a spinal cord neuron (*top right*). A change in voltage causes calcium channels to open, allowing an influx of calcium ions. This action initiates a series of molecular events that prompt neurotransmitters to be released from vesicles in the peripheral neuron. The neurotransmitters cross the synapse and dock with receptors that trigger the sending of a signal up the spinal cord to the brain. The drug Prialt blocks the calcium channels (*bottom right*) and thus the release of the neurotransmitters that convey the pain signal.

LUCY READING-IKKANDA

serve as an analgesic because of new findings showing that one way morphine works is by shutting down indirectly the activity of a specific subtype (the *N*-type) of calcium channel. Radiolabel and electrophysiological testing of SNX-111 demonstrated that it binds selectively to the *N*-type calcium channel. This obstructing action prevents the channel from opening and allowing an inrush of calcium ions. The ions precipitate transmission of a stimulus across the synapse into the spinal cord and then into the brain, where it is perceived as pain. "This was the smoking gun that allowed us to move forward," Miljanich comments. Subsequent animal tests indicated that the drug candidate exerted its effects with a thousandth the dose of opiates and without creating tolerance or addiction.

Nature's Gift

IN 1995 NEUREX STARTED a clinical trial of SNX-111 for patients with severe pain who had not been aided by intrathecally delivered opiates—those channeled by catheter from an implanted pump into the spinal fluid. The same pump delivered the synthetic snail peptide to select patients, restricted by the FDA to only those who were terminally ill, because of the previous side effects in the earlier trial. Following years of fiddling with the peptide's amino acid sequence, the drug that patients received in both clinical trials was the precise synthetic copy of the peptide found in the snail. "After testing hundreds of analogues, we came back to what nature gave us," Miljanich remarks.

Once the pain trial began, it became clear very quickly that the dosing regimen was wrong. Some patients encountered severe adverse side effects, ranging from a lack of coordination to auditory and visual hallucinations. Later trials specified that doses be reduced, and the period over which they were stepped up was greatly extended. In some cases, the side effects diminished. But not always. A patient's delirium, in one instance, ended only after electroconvulsive therapy. When approving the drug, the FDA specified that patients should be monitored frequently for neuropsychiatric symptoms.

In 1998 drugmaker Elan initiated a takeover, and Neurex quickly embraced the bid. Even though the results of two late-stage clinical trials demonstrated that SNX-111 provided significantly more pain relief than a placebo, Neurex officials knew that the FDA would probably ask it for another clinical trial before approval because of the drug's history of side effects. A Neurex late-stage study to protect brain cells from trauma had been put on hold until 2000 because of unimpressive results from earlier studies. An unrelated hypertension drug that Neurex had licensed to bring in some revenue might not have been enough to tide the struggling firm over until it received approvals for its two principal drugs.

SNX-111, now Prialt, went through another successful clinical trial and weathered a financial debacle that resulted from Elan's dubious accounting practices involving a series of joint ventures. Elan received FDA endorsement for Prialt on December 28 for neuropathic and additional types of severe chronic pain that do not respond to other forms of treatment. Approved eight years or so after its original target date, Prialt may net Elan more than $150 million a year if it gets prescribed to roughly 20 percent of the 55,000 or so patients with intrathecal catheters, at an estimated patient cost of $15,000 a year. "It's pretty exciting, although there have been so many ups and downs, milestones met and not met, that this is just another milestone," Miljanich says. For physicians who treat those with chronic pain, the news may be more upbeat. "Prialt has relieved pain for patients in whom all other therapies have failed," says Michael Leong, a physician who has supervised clinical trials of the drug and also served as a paid consultant to Elan.

Prialt is most likely Elan's last venture into snail peptides. The company did not preserve its capability, inherited from Neurex, to research cone snail toxins. Yet cone snails may still have much to offer as nature's combinatorial chemists. An article by two Harvard University biologists that appeared in 1999 in the *Proceedings of the National Academy of Sciences USA* found that two species of cone snails have toxin genes that could be the fastest-evolving on earth, an adaptation to changing prey that inhabit tropical reef ecosystems.

A few companies are planning to exploit these riches. Olivera, a father of the field, is a founder of Cognetix in Salt Lake City. Two Australian biotech outfits—Xenome and Metabolic Pharmaceuticals—have begun development or actual trials on snail peptide–based drugs, primarily for chronic pain. In some cases, Prialt and other peptide drugs in the works may be overtaken by development efforts that use small organic molecules, delivered orally, to target calcium and other ion channels.

Still, even small-molecule designers may use the snail peptides as a starting point for formulating new drugs. As appreciation grows for cone snails' chemical creativity, their home countries may become more proprietary about these genetic gold mines. Philippine newspapers have at times raised the specter of biopiracy. And a 2003 letter to *Science* written by researchers from Harvard Medical School, York University and the University of Chicago Pritzker School of Medicine decried the danger to cone snails and their habitats from collectors, coastal development, pollution and climate change, among other causes. The letter estimated that hundreds of thousands of the animals may be processed yearly by U.S. researchers, although a response to the letter calculated that on average not more than 5,000 cone snails a year need to be sacrificed to extract toxins that are analyzed and then synthesized. One laboratory even maintains a cone farm, enabling the milking of the mollusks without killing them.

Bioweapons may be as much of a worry as biopiracy. Since September 11, 2001, scientists engaged in cone snail research have found that rules for working with the toxins have grown much harsher, even though some venoms do not harm humans. "We're always terrified that we're breaking rules that we don't know about," Olivera says. Some caution may be warranted. The small size of the peptides, which makes them easy to synthesize, has at times held an allure for manufacturers of biological weapons. Before President Boris Yeltsin ordered the shutting down of Russia's bioweapons program in 1992, investigators there were trying to insert the gene for a lethal *C. geographus* peptide into the genome of the smallpox virus, which would have delivered a devastating double punch to victims. The Russian team failed in fashioning this bizarre smallpox-conotoxin hybrid, which could have approached 100 percent lethality. Chillingly, they probably just ran out of time. "The problems could have been solved," says Serguei Popov, a former top bioweapons researcher who is now a biology professor at George Mason University.

The tale of the cone snail, a lowly mollusk that has ascended toward the top of the marine food chain, will continue to fascinate. In 1998, under the pen name of Paul Adirex, a prominent Thai politician wrote a book, *The King Kong Effect,* about a plot to assassinate an American president using cone snail venom. Biologists have no need to resort to fantasy. The 50,000-plus cone snail peptides will keep them pondering this evolutionary wonder for decades to come. SA

Genes for cone snail toxins may be the fastest-evolving on earth. Their diversity constitutes A POTENTIAL PHARMACOLOGICAL GOLD MINE.

LUCY READING-IKKANDA

MORE TO EXPLORE

A New Way to Spell Relief: V-e-n-o-m. W. Wayt Gibbs in *Scientific American,* Vol. 274, No. 2, pages 28–30; February 1996.

Secrets of the Killer Snails. Alisa Zapp Machalek. Findings, National Institute of General Medical Sciences, September 2002. Available at **www.nigms.nih.gov/news/findings/sept02/snails.html**

Ziconotide: Neuronal Calcium Channel Blocker for Treating Severe Chronic Pain. G. P. Miljanich in *Current Medicinal Chemistry,* Vol. 11, No. 23, pages 3029–3040; December 2004.

A prodigious resource for all things cone snail, maintained by Bruce Livett, professor of biochemistry and molecular biology at the University of Melbourne, can be found at the Cone Shells and Conotoxins site: **http://grimwade.biochem.unimelb.edu.au/cone/index1.html**

TENNIS BALLS

Uniform Variety

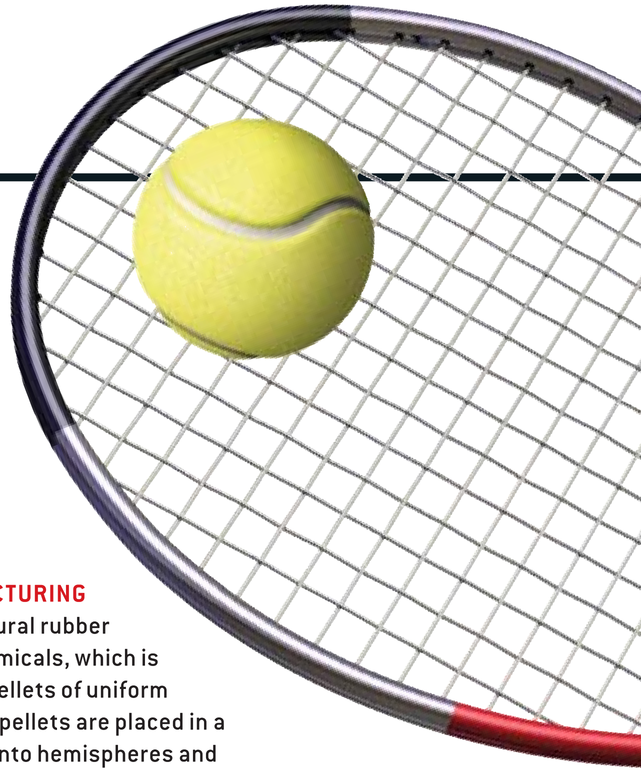
Manufacturers make more than 240 million tennis balls a year worldwide. And they are surprisingly uniform, given that they begin as natural rubber and wool, which vary with every barge and bale.

To be stamped "official" in accordance with the International Tennis Federation, a ball must meet rigid specifications for deformation and bounciness [*see table*]. Rubber, which "varies as much as lettuce," according to Lou Gagnon, technology manager at Head/Penn Racquet Sports in Phoenix, Ariz., is combined with up to 11 chemicals to create a homogeneous slurry. The mixture is pressed into molds to form the ball's center, or core. To craft a consistent cover, wool, nylon and cotton are woven into a felt that is soaked, shrunk and dried.

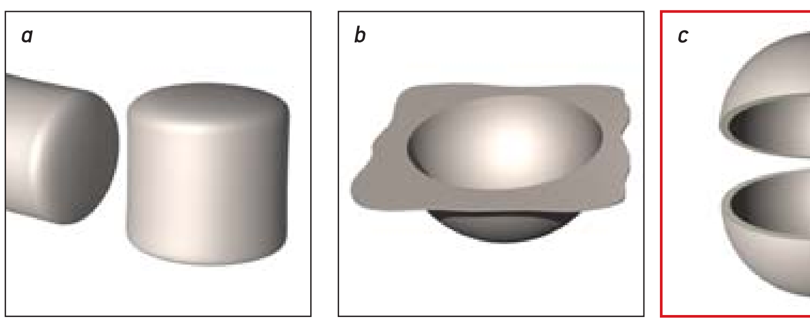
Ironically, in parallel with such efforts at consistency, the core's composition and the cover's nap length and tightness are fine-tuned to create one of three speed classes of balls: fast for slow courts like clay, medium for the ubiquitous hard courts, and slow for the fast grass surfaces. U.S. players prefer a fast game and generally play on hard surfaces, so the air inside balls is pressurized to about two atmospheres to make them more responsive. Many Europeans prefer slower play, typically on softer courts, so "nonpressurized" balls, sealed at one atmosphere, are also commonly sold.

The cover makes a big difference. "We use a different blend of fibers and a different [stiffening] procedure" for balls intended for clay versus hard courts, says Steve Judge, vice president at Tex Tech Industries in Tempe, Ariz. For example, a higher nap will be gripped better by racquet strings and will also increase aerodynamic drag. "Plus," Judge notes, "Europeans just prefer the look of a fuller ball."

Freshness counts, too. Depending on play, the air inside a pressurized ball will slowly force its way through the core and cover, making for a "flat" feel after so many months. To keep balls lively, manufacturers package them in canisters that are sealed at slightly less than two atmospheres of pressure. The near equilibrium prevents the air in the core from escaping—and it makes for a cool sound when the canister is popped open. —*Mark Fischetti*



BALL MANUFACTURING starts with natural rubber mixed with chemicals, which is extruded into pellets of uniform weight (*a*). The pellets are placed in a mold, pressed into hemispheres and heated (cured) at about 300 degrees Fahrenheit to attain a desired stiffness (*b*). The edges are buffed and coated with adhesive (*c*).



BALL HALVES are set closely opposite each other inside shrouds (*exploded view*). Air inside the shrouds is pressurized to two atmospheres. A press squeezes the shrouds together so the adhesive seals the halves and entraps the high-pressure air. The chamber is heated to cure the seal. The completed ball "center" is cooled with water and released (*d*).

GEORGE RETSECK

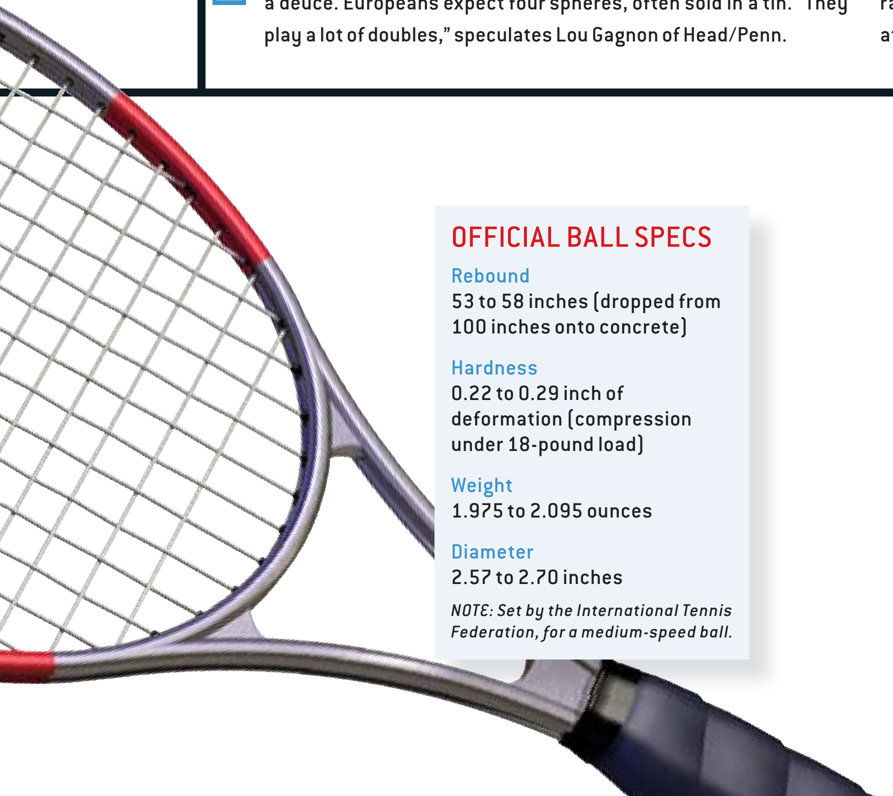
DID YOU KNOW . . .

BOUNCING AROUND: Natural rubber for tennis balls is vulcanized—chemically treated to set stiffness and durability—in a process similar to that for fabricating tires. British ball maker Dunlop Slazenger vulcanizes ball cores in the Philippines using clay from South Carolina, silica from Greece, sulfur from South Korea, zinc oxide from Thailand, and magnesium carbonate from Japan, among other compounds.

NOT JUST A THREESOME: Culture affects everything. Americans get three tennis balls when they buy a canister. But in Japan a tube holds a deuce. Europeans expect four spheres, often sold in a tin. "They play a lot of doubles," speculates Lou Gagnon of Head/Penn.

SEEING YELLOW: For decades, tennis balls were white. But makers shifted to optic yellow in the 1970s so that the balls would be easier to spot on television against the backgrounds of green courts, blue skies, players' white apparel, and the rainbow of spectators' garments.

PICKY PROS: More than 70,000 balls are delivered to each annual U.S. Open tournament. Balls are used for only nine games of a match before they are replaced. A player who faults on the first serve will often search for a ball with a puffier nap that will grip the racquet better, helping to impart spin and ensure that the second attempt drops in. Used balls go to charity.



OFFICIAL BALL SPECS

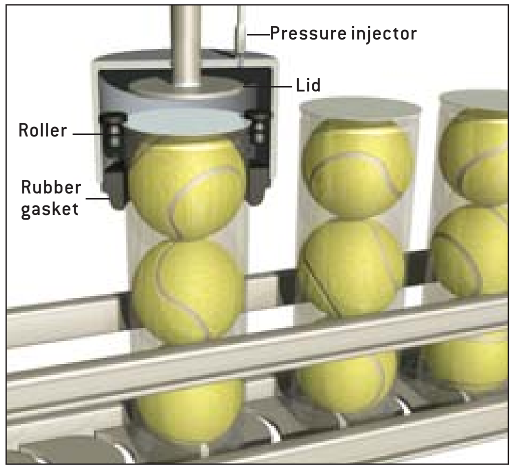
Rebound
53 to 58 inches (dropped from 100 inches onto concrete)

Hardness
0.22 to 0.29 inch of deformation (compression under 18-pound load)

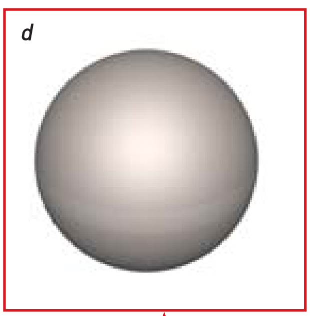
Weight
1.975 to 2.095 ounces

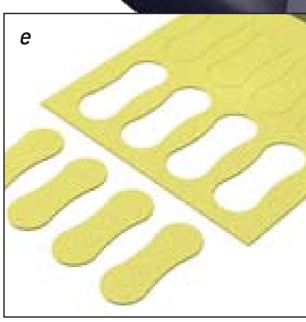
Diameter
2.57 to 2.70 inches

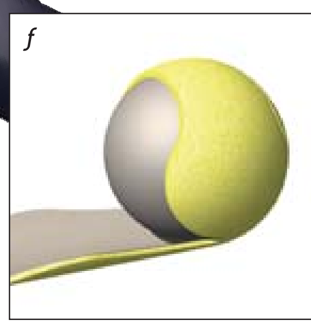
NOTE: Set by the International Tennis Federation, for a medium-speed ball.

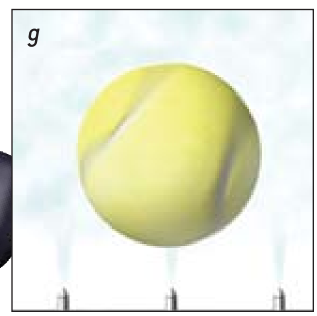


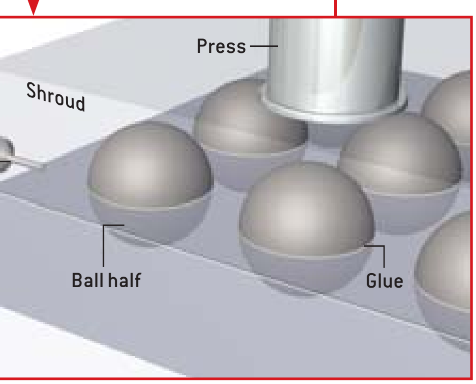
OPEN CANISTER of balls is raised into a chamber. Pressurized air is injected as an aluminum lid is lowered. Rollers then crimp the metal and plastic into an airtight seam.











COVER begins as wool is blended with nylon and spun into yarn. Next it is woven with a cotton yarn that forms the backing. The cloth is then soaked and passed through rollers that shrink the fibers, creating a felted surface. The felt is dried under tension to give it a desired stiffness. Large rolls of felt are coated on the back side with heat-activated adhesive. Two "dog bone" shapes are cut for each ball (*e*) and are pressed onto the ball center and cured (*f*). Finished balls are spun inside a tumbler (*g*), and steam is injected to raise the felt's nap.

Topic suggested by reader Myron Kayton.
Send ideas to workingknowledge@sciam.com

Hot Stuff

NEW THERMAL CAMERAS SHOW THE WORLD IN INFRARED BY MARK ALPERT

If, like me, you were an avid reader of comic books during the 60s and 70s, you'll probably remember a small advertisement that often ran on one of the back pages. Under the words "X-RAY GOGS," the ad showed a boy wearing a pair of glasses with lightning bolts radiating from the lenses to the boy's upraised hand, which looked like a skeleton's. The text below the picture asked, "Ever seen the bones in your hand?" For a nine-year-old, this come-on was irresistible, and I would've certainly mailed in $1.98 (plus postage) to purchase the goggles if my parents hadn't stopped me. "Don't waste your money," my father said.

WHAT'S HOT? The Thermal-Eye 250D, an infrared camera designed for police officers, can detect the body heat of fugitives and lost hikers.

Three decades later the lure of x-ray vision is still strong. A mini furor erupted a few years ago when photographers discovered that placing lens filters on certain Sony camcorders enabled them to see through thin skirts and swimsuits. The secret wasn't x-rays but infrared radiation, which we perceive as heat. The CCD (charge-coupled device) chips in the video cameras can detect some infrared rays as well as visible light; if you filter out the visible wavelengths, the images show the radiant heat from the sun reflected off objects.

But the alarmist stories that appeared on television ("A Camera That Can See Through Clothes!") greatly exaggerated the threat of thermal voyeurism. Only sheer fabrics are transparent to infrared, and the camcorders can observe just the near-infrared part of the spectrum—wavelengths shorter than one micron—so they cannot detect body heat, which peaks in intensity at about nine microns.

Until recently, true thermal cameras—those that can sense the long-wavelength infrared rays—were bulky and very expensive. The standard instrument was a photon detector that had to be cooled to cryogenic temperatures, usually by liquid nitrogen. (At room temperature, the signal would be swamped by thermal "noise.") Heat-seeking missiles use these systems for targeting, and astronomers employ them to peer into the dusty hearts of galaxies. Over the past 10 years, however, manufacturers have developed cheaper and more portable infrared cameras that do not have to be cooled. Instead of detecting photons, these devices focus the incoming infrared radiation on heat-sensitive materials; the resulting changes in the physical properties of the material are then used to generate an image.

When I heard about this new technology, I felt like I was nine years old again. Now I could finally get a glimpse of superhuman vision! One of the leading manufacturers of uncooled thermal cameras, L-3 Communications Infrared Products in Dallas, Tex., let me evaluate the Thermal-Eye 250D, a model designed to meet the needs of law-enforcement agencies. Because the unit can observe infrared rays with wavelengths between seven and 14 microns, it is perfect for detecting body heat. Police officers can use the 250D as a night-vision scope, scanning dark alleys for fugitives or searching the woods for lost hikers. Thermal cameras have an advantage

JOHNNY JOHNSON

over conventional night-vision scopes, which show greenish images and are widely used by the military: whereas the conventional scopes simply intensify the available visible light, the 250D can operate in total darkness.

You can just imagine how delighted I was when the unit arrived at my office. (I didn't even have to pay for postage.) At first glance, the 250D resembles a run-of-the-mill video camera. It has a rechargeable battery, an adjustable eyepiece and a pair of focus buttons. In fact, you won't find any major differences from an ordinary camcorder until you remove the lens cover and behold the oversize circle of polished germanium. Unlike glass, germanium is transparent to infrared.

The 250D is a pyroelectric detector—it focuses infrared rays on barium strontium titanate (BST), a ceramiclike material that acts like a capacitor. When connected to a circuit, the material undergoes electric polarization: one end becomes positive, the other negative. Changes in the temperature of BST alter the degree of polarization, inducing a displacement current in the circuit. The BST is divided into a rectangular array; by measuring the displacement current for each section of the array, the detector creates a two-dimensional image showing the intensity of the incoming radiation.

I turned on the camera, squinted into the eyepiece and surveyed my office. I discovered that the lens is optimized for long-range use—I couldn't focus on anything closer than 10 feet away. But by backing up a bit and adjusting the controls, I was able to get a nice infrared picture of my work space. The image as a whole was bluish and grainy, but the hot spots stood out well. The rear of my computer monitor put out a warm glow, and the vent of the radiator flared like a blast furnace. The window, though, was solid black (it was quite cold outside at the time).

NIGHTTIME SCENE viewed by the naked eye (*top*) reveals little. But the same scene glimpsed through an infrared camera (*bottom*) clearly shows a struggle with a purse snatcher.

COURTESY OF L-3 COMMUNICATIONS INFRARED PRODUCTS

Next, I wandered the hallways of *Scientific American*, noting the brilliant heat signatures of all the printers and copiers that were turned on. I was a bit leery of training the camera on my colleagues; although the officials at L-3 Infrared Products had assured me that the 250D couldn't see through clothes, I didn't want to take any chances. In the end, I focused the camera on our editor in chief, John Rennie. Luckily for both of us, I saw nothing too revealing. The infrared image delineated the folds of his shirt, but that was about it. His face, though, showed interesting variations—his nose was black, much colder than his cheeks and forehead. Also, his fingers were darker than his palms. Maybe it wasn't as exciting as seeing the bones in your hand, but it was still pretty neat.

John and I played a few tricks with the camera. John placed his hand on the door to his office for about 10 seconds; after he removed it, I could see a handprint of heat on the metal. Then we decided to see if the camera really worked in complete darkness. We went into an office without windows, shut the door and turned off the lights. It was pitch-black, but when I looked in the 250D and scanned the room, I found John hiding in the corner, as bright as a neon sign.

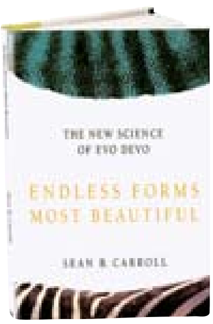
Finally, I took the camera to the famous skating rink at Manhattan's Rockefeller Center. Because the 250D looks so much like an ordinary camcorder, I figured I could pose as a tourist and unobtrusively observe the holiday crowds. In infrared, the scene was like something out of Dante's *Inferno:* a swarm of ghastly figures with glow-paint faces. In the background, the cars streamed down 49th Street, their front ends lit up by engine heat.

Whereas the 250D was designed for police work, other thermal cameras have been developed for firefighters, who use the devices to locate people trapped in dark, smoky buildings. L-3 Infrared Products has introduced a newer generation of detectors based on microbolometer technology; the devices focus infrared rays on an amorphous silicon material whose electrical resistance drops as its temperature rises. These units require less power than the 250D—they can run on a pair of AA batteries—and can fit in the palm of your hand. The Thermal-Eye X100xp, for example, weighs just 13 ounces. The U.S. Army has bought thousands of these devices, which are now in use in Iraq and Afghanistan.

The costs of the cameras are dropping as the technology advances, but they still run at least several thousand dollars apiece if you are buying them retail. The steep prices make me nostalgic for the $1.98 x-ray goggles. I'm still hoping to find out if those glasses really work. My kids will be old enough to read comic books soon, and if they're anything like me, I'm sure they'll be clamoring for a pair. SA

Evo Devo Is the New Buzzword …

… FOR THE 200-YEAR-OLD SEARCH FOR LINKS BETWEEN EMBRYOS AND EVOLUTION **BY BRIAN K. HALL**



ENDLESS FORMS MOST BEAUTIFUL: THE NEW SCIENCE OF EVO DEVO AND THE MAKING OF THE ANIMAL KINGDOM
by Sean B. Carroll
W. W. Norton, 2005 ($25.95)

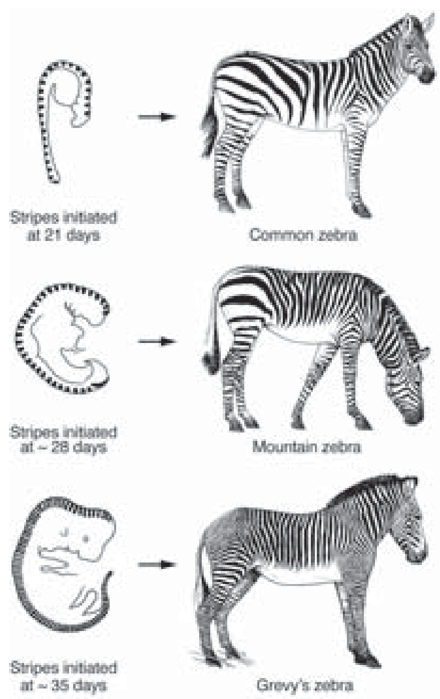
It would be hard to imagine two more different timescales in the lives of organisms than development—the transformation of an embryo to an adult within a single generation—and evolution—the modification and transformation of organisms between generations that reach back 600 million years. Yet for the past two centuries, natural philosophers, morphologists and biologists have asked whether there is a fundamental relationship between development (ontogeny) and evolution (phylogeny). There is, and it finds expression in the thriving discipline of evolutionary developmental biology (evo devo, as it has been called since the early 1990s).

Endless Forms Most Beautiful examines one of the most exciting aspects of evo devo—the incorporation of molecular biology that followed the discovery of classes of conserved regulatory (developmental, or "switching") genes: the homeobox, or *Hox,* genes. Carroll, who is a professor of genetics at the University of Wisconsin–Madison, writes in a lively style, peppering the book with endlessly fascinating examples that are beautifully illustrated by color and black-and-white drawings and photographs. To appreciate where this latest book devoted to evo devo is situated in the long history of the discipline, we need to go back almost 200 years.

The study of embryonic stages across the animal kingdom—comparative embryology—flourished from 1830 on. Consequently, when *On the Origin of Species* appeared in 1859, Charles Darwin knew that the embryos of all invertebrates (worms, sea urchins, lobsters) and vertebrates (fish, serpents, birds, mammals) share embryonic stages so similar (which is to say, so conserved throughout evolution) that the same names can be given to equivalent stages in different organisms. Darwin also knew that early embryonic development is based on similar layers of cells and similar patterns of cell movement that generate the forms of embryos and of their organ systems. He embraced this community of embryonic development. Indeed, it could be argued that evo devo (then known as evolutionary embryology) was born when Darwin concluded that the study of embryos would provide the best evidence for evolution.

Darwin's perception was given a theoretical basis and evo devo its first theory when Ernst Haeckel proposed that because ontogeny (development) recapitulates phylogeny (evolutionary history), evolution could be studied in embryos. Technological advances in histological sectioning and staining made simultaneously in the 1860s and 1870s enabled biologists to compare the embryos of different organisms. Though false in its strictest form, Haeckel's theory lured most morphologists into abandoning the study of adult organisms in favor of embryos—literally to seek evolution in embryos. History does repeat itself; 100 years later a theory of how the body plan of a fruit fly is established, coupled with technological advances, ushered in the molecular phase of evo devo evaluated by Carroll.

As Carroll discusses in his book (the



DIFFERENT NUMBERS OF STRIPES in three zebra species may result from differences in the time at which the stripes are initiated in the embryo. From the book: Drawing by Leanne Olds; modified from J. B. Bard in *Journal of Zoology,* Vol. 183, page 527; 1977.

COURTESY OF W. W. NORTON

title of which comes from the last lines of *The Origin of Species*), the discovery of Mendelian genetics in 1900, and soon after of the gene as the unit of heredity, thrust a wedge between development and evolution. Genes were now what mattered in evolution; embryos were merely the vehicles that carried genes from one generation to the next. Embryology was divorced from evolution, devo from evo. Even the discovery in the 1950s of the nature and role of DNA did not bring them back together. In the late 1970s, however, all began to change as several revolutions in theory and technology produced a mind shift as dramatic as the one that followed Darwin's *The Origin of Species*.

New methods for generating phylogenetic relationships brought comparative embryology back to the forefront; now we can assess the direction of evolutionary changes in development. When we find a species of frog that has lost the tadpole stage from its life cycle—a remarkable evolutionary change in form and function—we can determine whether that loss was an early or late event in the evolution of frogs. Stephen Jay Gould's seminal book *Ontogeny and Phylogeny* (1977) rekindled interest in 19th-century evolutionary embryology and resurrected an old idea—heterochrony, change in the timing of development in a descendant relative to an ancestor—in a form that could be tested. Important as these developments were, they were carried out against the then current wisdom that organisms differ because they possess unique genes not found in other organisms—lobster genes for lobsters, human genes for humans, and so forth.

The discovery of homeobox genes turned this approach upside down and inside out. The body plans of lobsters and humans, flies and fish, barnacles and mice, are initiated using the same families of genes that are conserved across the animal kingdom. The consequences of this discovery are the stuff of the first half of *Endless Forms Most Beautiful*, in which Carroll presents homeobox genes as the switches that contain the fundamental information required to make a fly's eye or a human hand.

The second half of the book explores what Carroll calls "the making of animal diversity," beginning with animal life as exemplified in the justly famous 500-million-year-old fossils of the Burgess Shale formation in British Columbia. Carroll is concerned with evolutionary tinkering with genetic switches and the production of patterns in nature—spots on butterfly wings, stripes on zebras. He devotes less attention to the downstream gene cascades and gene networks that allow similar signaling genes to initiate, for example, the wing of a bird or a human arm. Nor are the cells and cellular processes from which the endless forms are constructed given prominence.

Consequently, statements such as "the anatomy of animal bodies is really encoded and built ... by constellations of switches distributed all over the genome" could be taken to mean that switching genes contain all the information required to generate form. Were that true there would be no need for evo devo; indeed, there would be no development. It would all be geno evo. But, as Carroll demonstrates, "the evolution of form occurs through changes in development," which is precisely why evo devo is so central to understanding how animals have been and are being evolved. SA

Brian K. Hall is George S. Campbell Professor of Biology and a University Research Professor at Dalhousie University in Halifax. He is author of Evolutionary Developmental Biology *and of* Bones and Cartilage: Developmental and Evolutionary Skeletal Biology, *among other books, and co-editor with Benedikt Hallgrímsson of the forthcoming* Variation: A Central Concept in Biology.

THE EDITORS **RECOMMEND**

MORE THAN HUMAN: EMBRACING THE PROMISE OF BIOLOGICAL ENHANCEMENT
by Ramez Naam. Broadway Books, 2005 ($24.95)

CITIZEN CYBORG: WHY DEMOCRATIC SOCIETIES MUST RESPOND TO THE REDESIGNED HUMAN OF THE FUTURE
by James Hughes. Westview Press, 2004 ($26.95)

GOLEMS AMONG US: HOW A JEWISH LEGEND CAN HELP US NAVIGATE THE BIOTECH CENTURY
by Byron L. Sherwin. Ivan R. Dee, 2004 ($26)

These books consider, each from a different perspective, biotechnology as a major issue of the 21st century. Naam, who works on Internet search technology at Microsoft, argues that society should embrace "techniques that might enhance human abilities." Describing several of them, he says that people who would ban them "had better have strong evidence that the research poses a greater threat to society than the medical benefits it brings."

Hughes, who teaches health policy at Trinity College in Hartford, focuses on the political battles over cutting-edge biotech: "Transhuman technologies, technologies that push the boundaries of humanness, can radically improve our quality of life ... we have a fundamental right to use them to control our bodies and minds." But "we need to democratically regulate these technologies and make them equally available in free societies."

Sherwin, distinguished service professor of Jewish philosophy and mysticism at the Spertus Institute of Jewish Studies in Chicago, sounds a cautionary note. Technological achievement "not guided by a moral compass is ultimately self-defeating and self-destructive," he writes. "To devalue human life in an effort to enhance it is a pyrrhic victory, the kind epitomized by the exploits of Victor Frankenstein."

The books reviewed are available for purchase through www.sciam.com

Eye off the Ball

CAN FOOTBALL'S ANCILLARY ANTICS POSSIBLY OUT-INSULT THE GAME ITSELF? **BY STEVE MIRSKY**

Outlined against a blue-gray October sky the Four Horsemen rode again. In dramatic lore they are known as Famine, Pestilence, Destruction and Death. These are only aliases. Their real names are Concussion, Arthritis, Torn Anterior Cruciate Ligament and The Pile.

It's football season—for me, anyway. True, this is the April issue, which coincides with the start of the baseball season, thank goodness. But I'm writing these words in the week leading up to the Super Bowl, so football is on my mind—specifically the way the game chews up and spits out the bodies of its participants, willing though they be.

There are two reasons I've been thinking about football and its injuries—Janet Jackson and Randy Moss. The Jackson "wardrobe malfunction" at the 2004 Super Bowl is now legendary. (I didn't see it myself, because anyone actually watching a Super Bowl halftime show has either lost the remote or is researching an anthropology doctoral thesis.) Congress rapidly organized hearings on Jackson's televised breast, clearly the greatest threat faced by our nation in a presidential election year. Irate congresswoman Heather Wilson of New Mexico quoted her young son, who allegedly called the incident "nasty."

Then, early in this season's pro football playoffs, Minnesota Vikings wide receiver Randy Moss taunted the Green Bay Packers crowd—he made-believe he was pulling down his pants and then wagged his booty at the fans. Announcer Joe Buck was especially livid, calling Moss's pantsomime "disgusting."

I recap the Jackson and Moss incidents that so outraged so many so that I can ask: Have you ever watched the actual game?

Force equals mass times acceleration, and football features a lot of massive guys violently accelerating into one another with devastating force. Researchers look at the effect of these forces at the Center for the Study of Retired Athletes, a joint (perhaps the only intact joint in football) project of the University of North Carolina at Chapel Hill and the National Football League Players Association. In an attempt to quantify the toll the game takes on its professional participants, the center surveyed more than 2,500 former players in 2001.

So here's what the game does to its players during the family-friendly part of the telecast. Of survey respondents, 62.5 percent had gotten at least one concussion. The average was two concussions. A quarter suffered three or more concussions. This last group has three times the normal risk of depression. They also have an elevated risk of the kind of cognitive impairment that often precedes full-blown Alzheimer's disease. Moving from neurology to orthopedics, 38 percent of ex-players have osteodegenerative arthritis. Former players between the ages of 25 and 54 have much higher rates of arthritis than regular guys the same ages. Twenty percent of players damaged their knees' anterior cruciate ligament, which can lead to permanent mobility problems.

Football players whose games are on artificial turf face an additional danger—rug burns. And burned players get antibiotic treatment at 10 times the rate of guys who don't get mauled on the carpet. All those drugs can lead to resistance—a study published in the February 3 *New England Journal of Medicine* found that in 2003 five St. Louis Rams who played on fake grass at home wound up with drug-resistant skin staph infections.

Further evidence of football's wholesome nature can be found in the January 31 issue of *Sports Illustrated,* which featured trenchant testimony by current pro players about what goes on inside a pile, where everyone is battling everyone else for the ball: "I've had guys go for the privates, guys try to put their elbow in my neck, guys reaching inside my helmet." "The go-to spots are the eyes and the family jewels." "Guys reach inside the face mask to gouge your eyes. But the biggest thing is the grabbing of the testicles. It is crazy." Indeed, it's nuts in there. But at least it's not nasty or disgusting. SA

ILLUSTRATION BY MATT COLLINS; PHOTOGRAPH BY FRANK VERONSKY

ASK THE EXPERTS

What is the fastest event that can be measured?

—R. Mitchell, Melbourne, Australia

Scott A. Diddams and Thomas R. O'Brian of the Time and Frequency Division of the National Institute of Standards and Technology offer an explanation:

The answer depends on how one interprets the word "measured." Both the accurate measurement of fleeting events and the recording or inference of such occurrences are of interest. So we suggest rephrasing the original query into two new ones: "What are the shortest time spans that can be measured with a particular accuracy?" and "What are the briefest happenings that can be recorded or inferred?"

Currently cesium atomic-fountain clocks are the best way to measure time with a certain accuracy. These "clocks" are actually frequency standards rather than timekeeping devices, and they achieve the defined cesium standard frequency with the exceptional precision of about one part in 10^{15}. Put another way, in 30 million years of continuous operation, they would neither gain nor lose more than a second. Yet these frequency standards are rather "noisy," and to achieve such impressive results requires averaging many thousands of separate frequency measurements over a period of about one day. So fountain clocks are not generally useful for timing short-duration events.

Among the shortest-period events that can be directly created, controlled and measured are quick bursts of laser light. These pulses occur on timescales of femtoseconds (10^{-15} second) and, more recently, attoseconds (10^{-18} second). Femtosecond pulses can function in a manner similar to the flash on a camera used to "freeze" events that are too fast for the eye to register. A source of femtosecond-order light pulses is a mode-locked laser, in which many optical waves cooperate to produce a pulse. It is yet another problem, however, to accurately measure that pulse's duration. No photodetectors or electronics are fast enough, so scientists commonly employ correlation techniques that translate a temporal measurement into a distance measurement. To date, pulses as short as a few hundred attoseconds have been generated and measured.

Now let us consider the second question, regarding the most transitory episodes that we can record or infer. Events as short as about 10^{-25} second have been indirectly inferred in extremely energetic collisions in the largest particle accelerators. For example, the mean lifetime of the top quark, the most massive elementary particle so far observed, has been inferred to be about 0.4 yoctosecond (0.4×10^{-24} second).

Why is normal blood pressure less than 120/80? Why don't these numbers change with height?

Jeffrey A. Cutler, senior scientific adviser for the National Heart, Lung, and Blood Institute at the National Institutes of Health, responds:

The origin of the designation of 120/80 as the threshold for "normal" blood pressure is unknown. (The top number is the systolic pressure, which is the pressure in the arteries while the heart is pumping; the bottom is the diastolic pressure, a measure of pressure in the arteries while the heart is resting and refilling with blood.) It may have come from data available early in the 20th century from life insurance physical exams. In any case, epidemiological studies confirm that the risk of a heart attack or stroke begins to increase in adults when the systolic is 115 or greater or the diastolic is 75 or more. The risk steadily increases with higher and higher readings, so the traditional 120/80 level remains a reasonable guideline for getting a doctor's attention, with the main goal of preventing your pressure from continuing to rise in subsequent years.

Blood pressure does in fact increase with height, ensuring that the brain, located at the uppermost point of the circulatory system for most of the day, gets sufficient blood flow and oxygen. But the effect is fairly small, which is why the 120/80 figure is not adjusted for taller people. SA

For a complete text of these and other answers from scientists in diverse fields, visit www.sciam.com/askexpert

MATT COLLINS