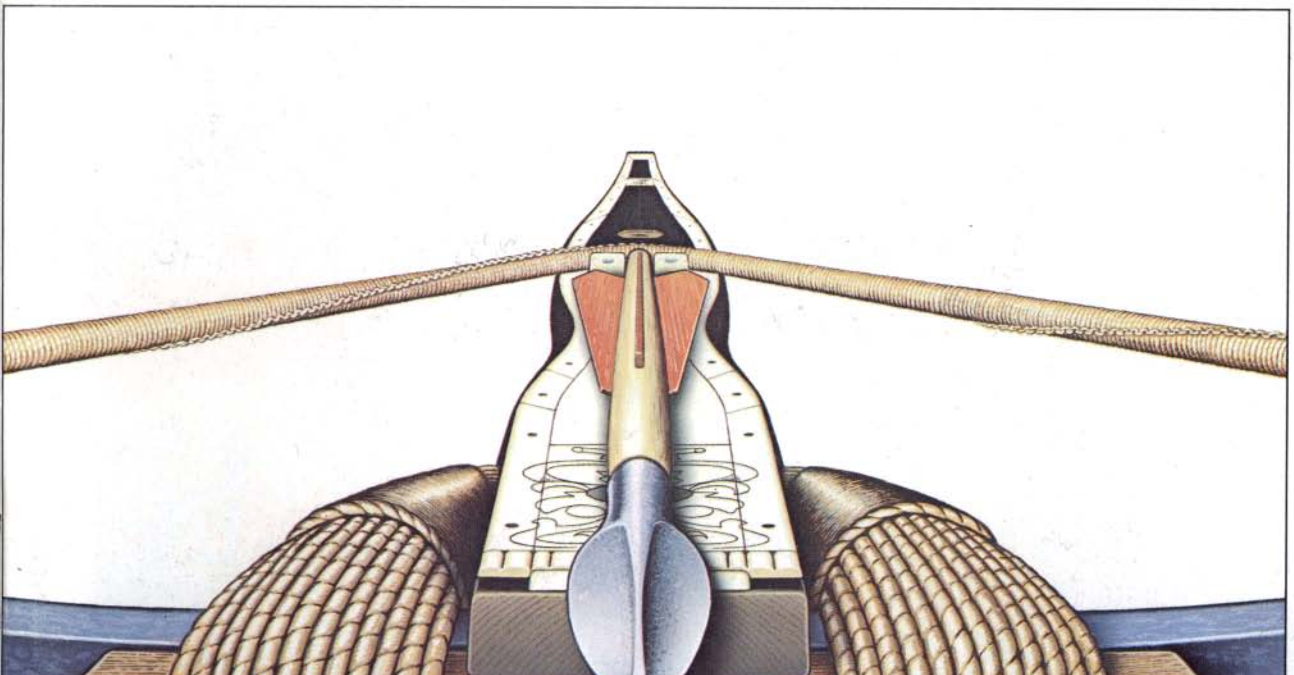
# SCIENTIFIC AMERICAN

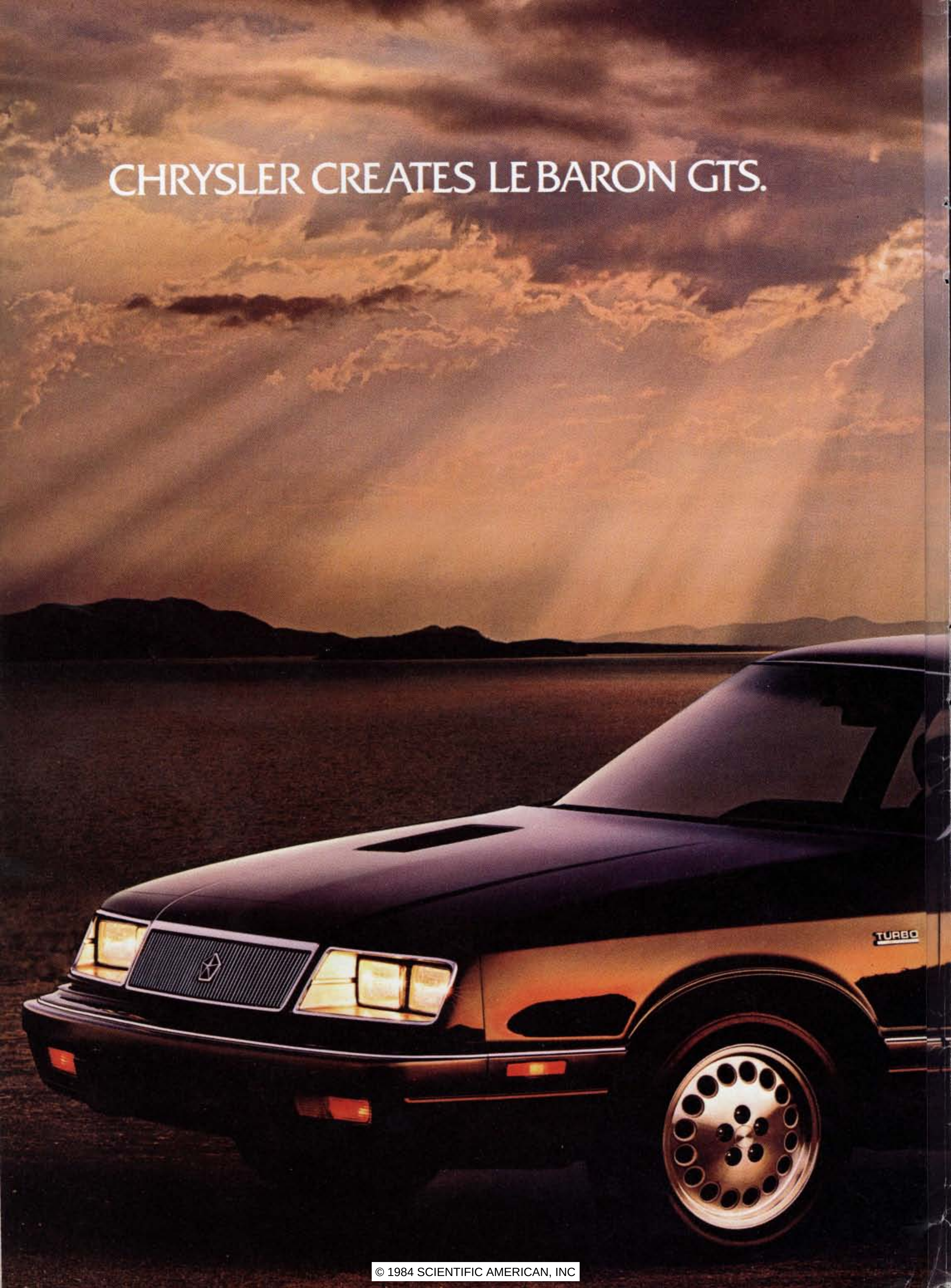


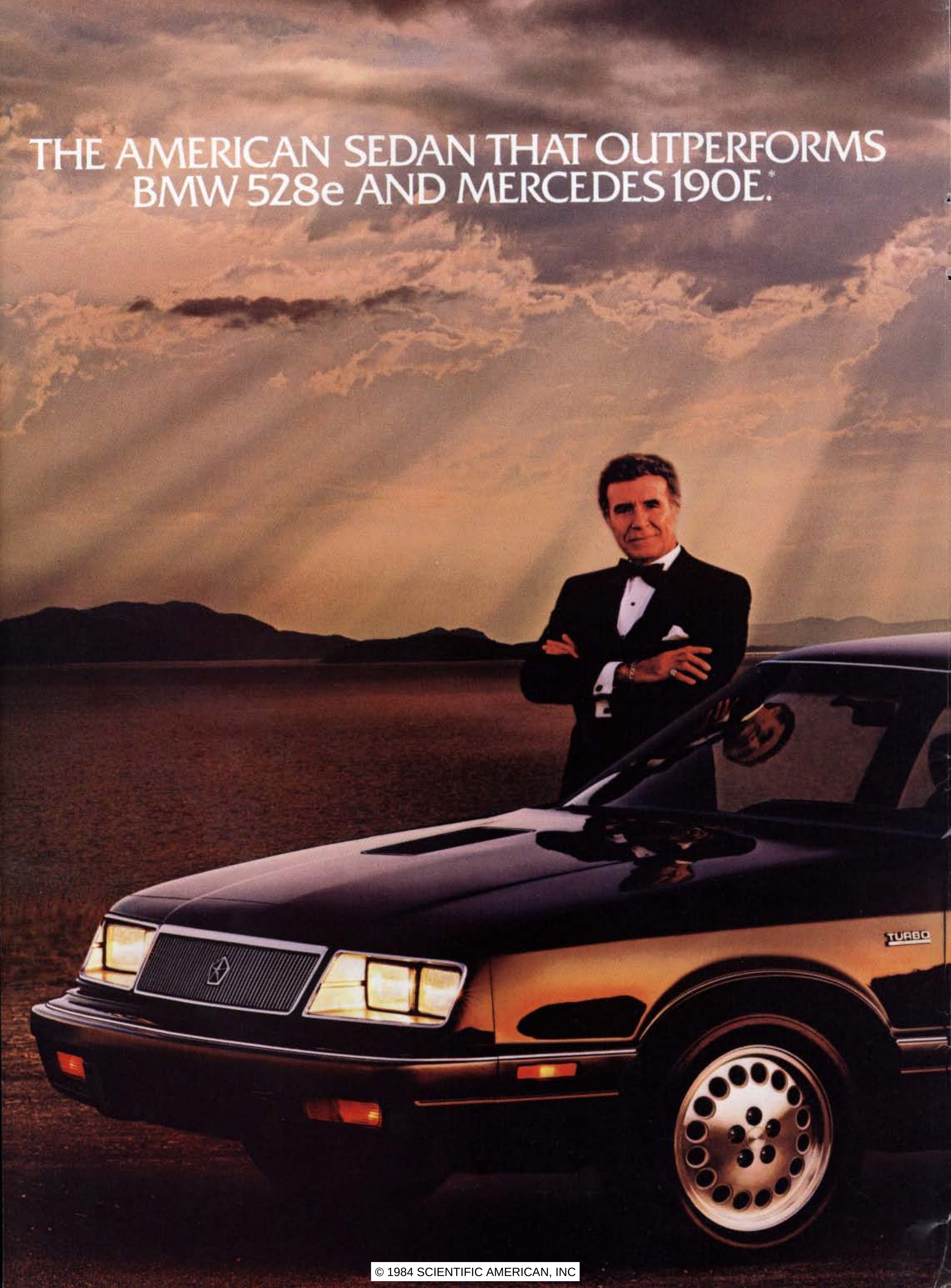
THE CROSSBOW

$2.50

January 1985


Dwight Knott,
Sun Company manager
of the Big Horn Ranch
and Reclamation
Research Center.


CHRYSLER CREATES LE BARON GTS.
TURBO


THE AMERICAN SEDAN THAT OUTPERFORMS
BMW 528e AND MERCEDES 190E.*
TURBO


RPM X 1000
MPH
UNLEADED FUEL ONLY



Established 1845 **SCIENTIFIC AMERICAN** January 1985 Volume 252 Number 1

ARTICLES



DEPARTMENTS





SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1984 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017

# SCIENTIFIC AMERICAN

## CORRESPONDENCE

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

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

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

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

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

Name ______________________

**New Address** ______________________

Street ______________________

City ______________________

State and ZIP ______________________

**Old Address** ______________________

Street ______________________

City ______________________

State and ZIP ______________________

### THE COVER

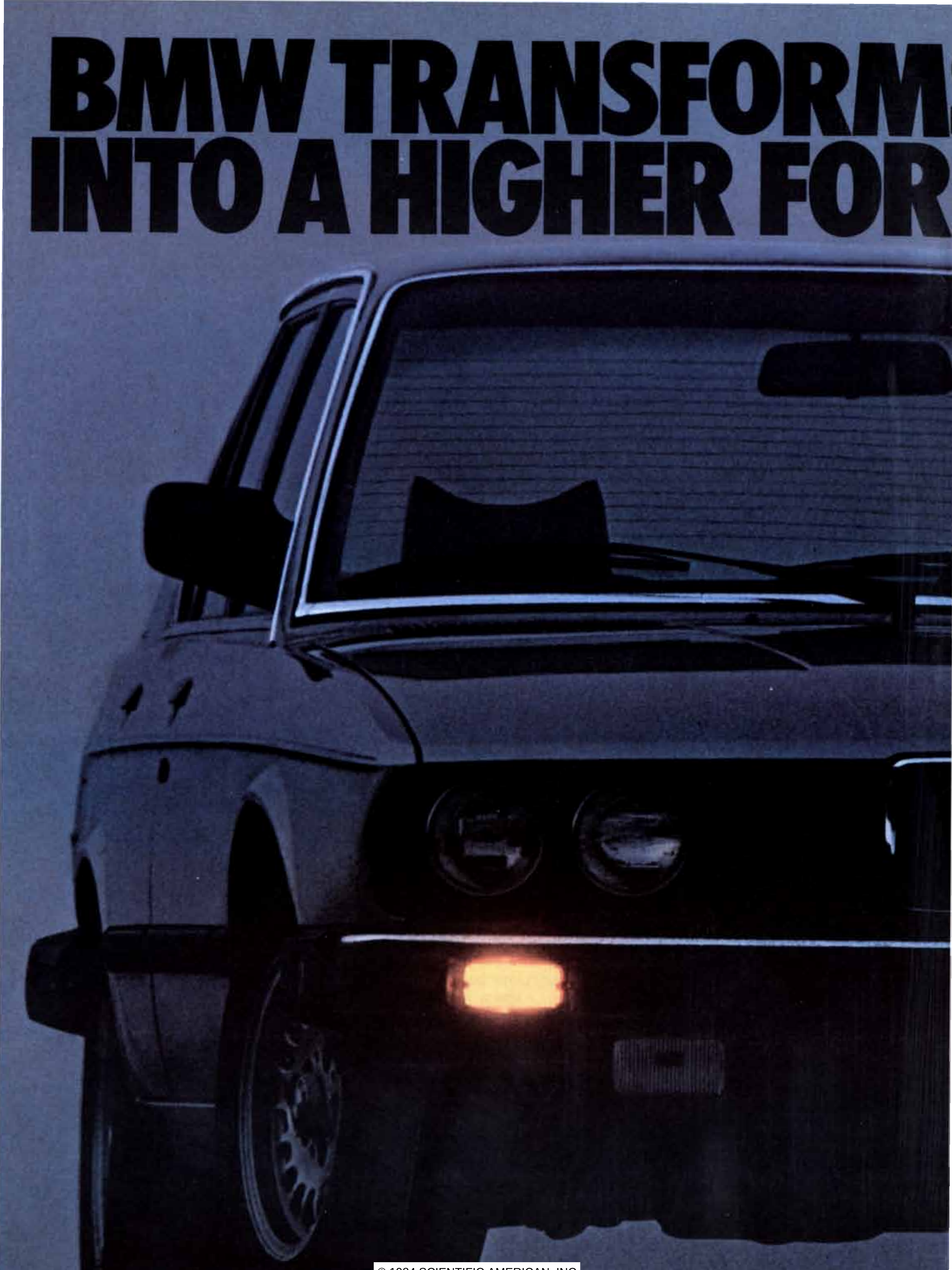
The painting on the cover depicts an English crossbow bearing the date 1617 on its black wood tiller, or stock. The inlaid ivory on the upper surface of the tiller suggests that the bow was made for hunting; a military crossbow would probably not have been so elaborately decorated (see "The Crossbow," by Vernard Foley, George Palmer and Werner Soedel, page 104). To cock the bow, which has a draw weight of several hundred pounds, the archer probably employed a cranklike, geared device called a cranequin; the tiller has holes that evidently served for attaching such a mechanism. The string is shown in the firing position. It is held there by a release mechanism that can be depressed by a pull on the trigger, which is on the bottom of the tiller. Thus fired, the 12-inch bolt, or arrow, would travel as far as 400 yards. The ring-and-rope arrangement in the foreground fastens the bow to the tiller. This bow is part of a collection of several crossbows maintained by the museum of the U.S. Military Academy at West Point, N.Y.

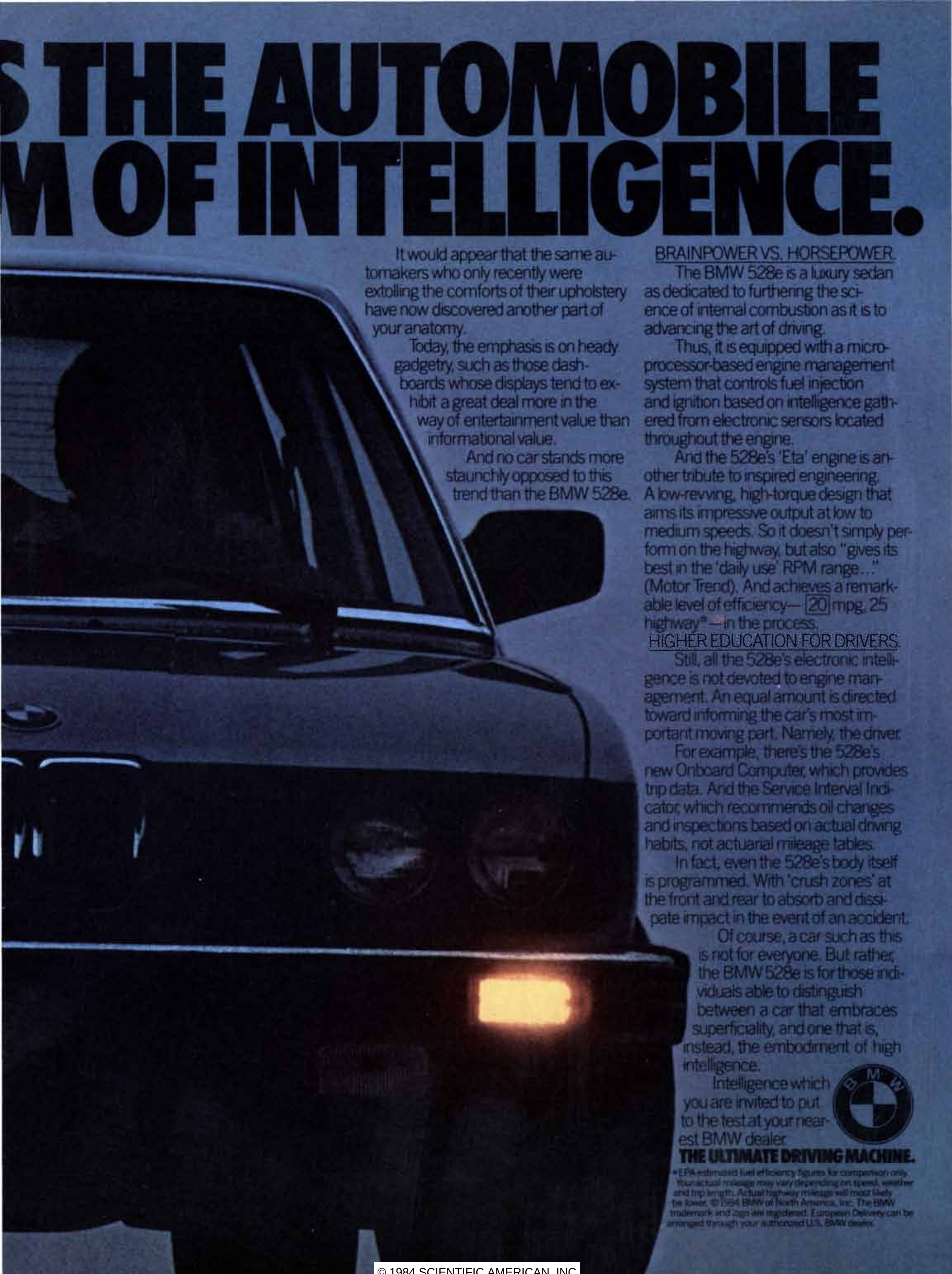
### THE ILLUSTRATIONS

Cover painting by Hank Iken

| Page | Source |
|---|---|
| 11 | SCIENTIFIC AMERICAN |
| 61 | Ilil Arbel |
| 33 | U.S. Air Force |
| 34–39 | Hank Iken, Walken Graphics |
| 41 | Ian Worpole |
| 42 | David F. Malin, Anglo-Australian Observatory |
| 43–45 | Ian Worpole |
| 47 | James Kilkelly (*top*), Ilil Arbel (*bottom*) |
| 48 | *Canadian Journal of Psychology* |
| 49 | Ilil Arbel |
| 50 | *Cognition* |
| 51 | Patricia K. Kuhl, University of Washington |
| 52 | Ilil Arbel |
| 54 | National Aeronautics and Space Administration |
| 55 | Judah Folkman, Harvard Medical School |
| 56 | Lawrence Livermore National Laboratories |
| 57 | The Liposome Company, Inc. |
| 63 | B. Leonard Holman and Thomas C. Hill |
| 64 | Daniel P. Perl |
| 65 | Dennis J. Selkoe (*top*), George G. Glenner (*bottom*) |
| 66–71 | Dennis J. Selkoe |

| Page | Source |
|---|---|
| 72 | Daniel P. Perl |
| 73 | Marek-Marsel Mesulam (*top*), Sally Black (*bottom*) |
| 74 | Sally Black |
| 77–78 | Douglas McWilliam, Lawrence Berkeley Laboratory |
| 79 | Alan D. Iselin |
| 80 | Bo Jakobsson, University of Lund (*top*); Alan D. Iselin (*middle*); Dana Beavis, University of California at Riverside (*bottom*) |
| 81–87 | Alan D. Iselin |
| 89 | Robert Paz, California Institute of Technology (*top*); Gabor Kiss (*bottom*) |
| 90–94 | Gabor Kiss |
| 97–101 | Patricia J. Wynne |
| 102 | John M. Gosline and M. Edwin DeMont, University of British Columbia |
| 103 | Patricia J. Wynne |
| 105 | The National Gallery, London |
| 106 | James Kilkelly |
| 107–109 | Andrew Christie after Susan Canganelli |
| 112–119 | Michael Goodman |


BMW TRANSFORM
INTO A HIGHER FOR



S THE AUTOMOBILE
M OF INTELLIGENCE.
It would appear that the same automakers who only recently were extolling the comforts of their upholstery have now discovered another part of your anatomy.
Today, the emphasis is on heady gadgetry, such as those dashboards whose displays tend to exhibit a great deal more in the way of entertainment value than informational value.
And no car stands more staunchly opposed to this trend than the BMW 528e.
BRAINPOWER VS. HORSEPOWER.
The BMW 528e is a luxury sedan as dedicated to furthering the science of internal combustion as it is to advancing the art of driving.
Thus, it is equipped with a microprocessor-based engine management system that controls fuel injection and ignition based on intelligence gathered from electronic sensors located throughout the engine.
And the 528e's 'Eta' engine is another tribute to inspired engineering. A low-revving, high-torque design that aims its impressive output at low to medium speeds. So it doesn't simply perform on the highway, but also "gives its best in the 'daily use' RPM range..." (Motor Trend). And achieves a remarkable level of efficiency— 20 mpg, 25 highway*—in the process.
HIGHER EDUCATION FOR DRIVERS.
Still, all the 528e's electronic intelligence is not devoted to engine management. An equal amount is directed toward informing the car's most important moving part. Namely, the driver.
For example, there's the 528e's new Onboard Computer, which provides trip data. And the Service Interval Indicator, which recommends oil changes and inspections based on actual driving habits, not actuarial mileage tables.
In fact, even the 528e's body itself is programmed. With 'crush zones' at the front and rear to absorb and dissipate impact in the event of an accident.
Of course, a car such as this is not for everyone. But rather, the BMW 528e is for those individuals able to distinguish between a car that embraces superficiality, and one that is, instead, the embodiment of high intelligence.
Intelligence which you are invited to put to the test at your nearest BMW dealer.
THE ULTIMATE DRIVING MACHINE.
*EPA estimated fuel efficiency figures for comparison only. Your actual mileage may vary depending on speed, weather and trip length. Actual highway mileage will most likely be lower. © 1984 BMW of North America, Inc. The BMW trademark and logo are registered. European Delivery can be arranged through your authorized U.S. BMW dealer.

Grand Marnier
Liqueur
Paradise found.
For gift delivery anywhere call 800-528-6148 (except where prohibited by law).
Product of France. Made with fine cognac brandy 80 proof. © 1985 Carillon Importers, Ltd., N.Y.

# LETTERS

Sirs:

In the item "Ambiguous Protection" ["Science and the Citizen," September, 1984] there is a description of a new idea for software protection that involves writing ambiguous bits onto a diskette. Would it not be better to create an area that would produce an input error when read? The program need only read this area once. If an input error does not occur, the program will assume that the diskette on which it resides is an unauthorized copy and will not run it.

What is needed to make my suggestion practical is one of the following two techniques:

1. Create "bad" bytes by a mechanical device.

2. Design the computer to create a parity bit after writing each block of information. The total number of bits in the information plus the parity bit would be odd or even depending on whether an odd or an even parity scheme was chosen. When asked to read the diskette, the hardware would verify that the parity is correct. A software manufacturer wanting to protect a diskette would use a special computer to write an area with the parity opposite to that used by the system for which the software is intended. Reading this area would produce an input error, which is the desired result.

SEYMOUR BLOOM

San Jose, Calif.

Sirs:

The techniques proposed by Mr. Bloom are already in use in certain commercial products. However, the parity method can be defeated easily on most microcomputers by tampering with the operating system, and the mechanical method is quite expensive since each diskette must be treated individually by high-precision tools. The ambiguous-bit technique, on the other hand, requires only a one-time expenditure on modifying the disk drives used for legitimate software duplication, and the cost per diskette becomes negligible in mass production.

ADI SHAMIR

Cambridge, Mass.

Sirs:

It is surprising to see *Scientific American* publish an article on the order of "The Climatic Effects of Nuclear War," by Richard P. Turco, Owen B. Toon, Thomas P. Ackerman, James B. Pollack and Carl Sagan [August, 1984]. Sagan has already treated us to his vision of a nuclear night via television. His politicized views were properly called speculative by S. Fred Singer on the editorial page of the *Wall Street Journal* (February 3, 1984).

It appears that Sagan and many others are afflicted with the problem of allowing their personal politics to get in front of their scientific work. Sagan's entire scenario falls apart on one very simple consideration. To this day atmospheric scientists and weather experts, using all the worldwide weather data available with the most modern theories and algorithms in the most powerful computers, cannot reliably and accurately predict the weather. The *Farmer's Almanac* is about as dependable as their best predictions. Yet the effect of nuclear explosions on local and worldwide atmospheric conditions must follow the same laws that control our everyday weather changes. If the weather cannot be predicted with present knowledge, neither can nuclear-night scenarios.

"The Climatic Effects of Nuclear War" is a highly speculative, perhaps scare-generating tract, not up to the standards of the usual *Scientific American* article.

CORNEL SAROSDY

San Antonio, Tex.

Sirs:

The unreliability of weather predictions is well known. Doubtless Mr. Sarosdy, like the rest of us, has been caught in a downpour without his overshoes. But weather is not the same as climate. It is much more difficult to predict correctly that it will rain on a given day in a given locale (weather) than it is to predict that it will be cold during wintertime (climate). We can calculate, using straightforward and reasonably elementary physics, the present climate of the earth, and in particular the surface temperature averaged over latitude, season and the diurnal cycle. The answer obtained is the correct, observed value for the earth's climate. While the computations for "nuclear winter" are considerably more difficult, they are much closer in methodology to calculations of the global climate in the current epoch than they are to predictions of the weather.

Of course, even the simplest climate-calculation techniques can be misunderstood or misapplied. Since the nuclear-winter problem involves several new phenomena, some scientists at first did not correctly understand the problem. Singer, for example, erroneously expected that a significant greenhouse effect could occur. Deficiencies in Singer's article were pointed out in a subsequent issue of the *Wall Street Journal* (February 16, 1984).

The nuclear-winter idea originated from our scientific studies of planetary climates and of terrestrial events such as climatic changes due to volcanic eruptions and extinctions caused by asteroid impacts. Like all scientific findings, this one can be shown to be true or false by objective means, independent of political predispositions. It is not subject to experimental verification, at least not more than once, and most of us do not want to perform the experiment. However, each of the elements of the hypothesis can be further tested, for example by collecting data on the amount of burnable material in targeted cities, by studying the altitude reached by the plumes of massive fires or by using improved techniques to calculate the climatic changes following a nuclear war.

Nuclear winter is neither a scare tactic nor a political issue. Instead it is a fact that must be considered when making political decisions about nuclear arsenals and nuclear war. We believe further scientific study is needed to understand nuclear winter more clearly. Conceivably we may have overestimated its severity; however, it it seems more likely that we have *under*estimated both its severity and its duration—for example by choosing to ignore the 30,000 tactical nuclear weapons in the world, to minimize the effects of fire storms and buoyancy in injecting smoke into the nominal stratosphere and to neglect climatic feedback. These omissions were made partly to guarantee the conservative nature of our results.

We do not find discussion of nuclear winter frightening. What we do find frightening, however, is that humans have constructed a global arsenal of immense destructive power without carefully considering the consequences of its use, and that people such as Mr. Sarosdy do not want to hear that nuclear war might be devastating for our planet and our species.

RICHARD P. TURCO

OWEN B. TOON

THOMAS P. ACKERMAN

JAMES B. POLLACK

CARL SAGAN

# 50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1935: "Plastics came into existence just 66 years ago when John Wesley Hyatt, searching for a substitute for ivory for making billiard balls, invented Celluloid, a nitro-cellulose product. But it was not until 1907, when Dr. L. H. Baekeland discovered how to control the union of phenol, or carbolic acid, and formaldehyde to make the synthetic resin known as Bakelite, that one could speak of plastics and a plastics industry. During the past 20 years plastics production has risen steadily. Now, following a few years of very rapid growth, the industry has an annual output conservatively valued at 50,000,000 dollars. Viewing the present position of plastics, one might think they had reached a stage of complete exploitation. Widespread as is their use, the limit is not yet reached."

"Revolution born of the test tube stalks among the petroleum refineries. The domination of Pennsylvania oils is threatened. Low-grade crudes now supply values of lubrication thought impossible of production a year or so ago. By developing a solvent refining process using propane, western manufacturers promise rich development of California fields producing wax-bearing crudes. The challenge was to isolate the desirable paraffin-base constituents. Experimentation proved that if a paraffin-base oil were to be refined from California crudes, it would be necessary to use the wax-bearing crudes in order to secure a yield economically feasible. This brought up the problem of de-waxing. It was the discovery that propane could be used as a de-waxing agent that made possible the production of this new type of oil from California crudes."

"A small electric motor, powered solely from sunlight and able to run continuously as long as the rays of the sun fall on the light-sensitive surface, has recently been constructed by a Detroit manufacturer, J. Thos. Rhamstine. A battery of 20 small light-sensitive and power generating disks, connected together and directly to a small direct current permanent-field motor, turns the motor at a high rate of speed without the use of any source other than the effect of the sun's light rays on the disks. This motor at present has no practical value, for the amount of power produced is very small. But to the scientifically minded it has significance, since the possibilities of obtaining power directly from the sun have scarcely been touched."

"Although dozens of comets have been observed through observatory telescopes in the past few years, and many more as bright are expected in the near future, astronomers do not know when one brilliant enough to be conspicuous to the naked eye will appear. It is reasonable to expect 'one or more great comets sometime within the next 50 years, but whether one will come next week or next year or not in the next ten years no one can say.' So reports Dr. Robert G. Aitken, director of the University of California's Lick Observatory. During the 19th Century, he states, 'five comets of the first rank appeared and at least six others that were fairly brilliant.' One of these was Halley's, which returned in 1835. Halley's made its next visit in 1910. The comet is confidently expected again in about 1985."

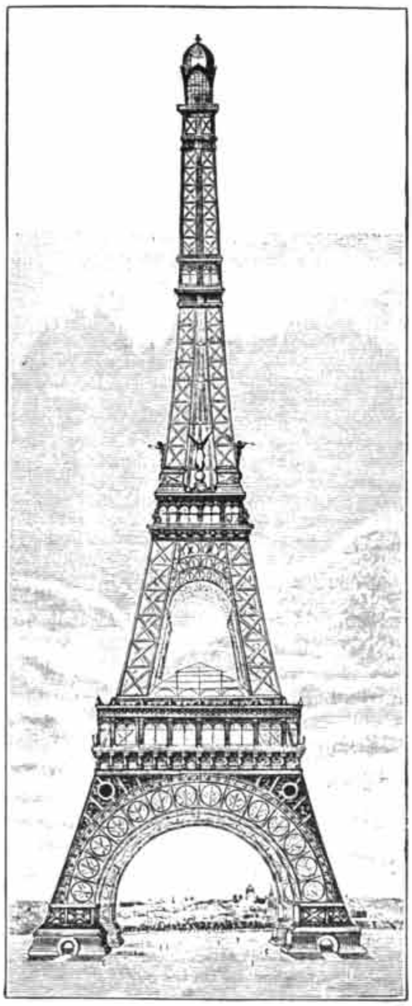
SCIENTIFIC AMERICAN

JANUARY, 1885: "The application of natural gas at Pittsburg and vicinity is watched with an abiding interest. With gas at fifteen cents per thousand feet, it is not strange that manufacturers should be eager to avail themselves of so cheap a fuel, so cleanly in use, so readily applied and controlled and, in many of its applications, furnishing a superior product. The present excitement in gas territory is not unlike the early oil days. Prospectors are sinking new wells, pipe line companies are organizing and law suits have been initiated, a sure sign of serious purpose, while millions of cubic feet are daily wasted. We predict for natural gas a future as great as that which grew out of the early petroleum development."

"Thirty-five years ago the name of St. Paul, Minn., was unknown to map makers, and neighboring Minneapolis could not boast of a single building. To-day each of these cities has a population little short of 100,000 souls, and each is the center of a robust growth in trade, manufactures and intellectual activity. There were buildings erected in the twin cities during the present year at a cost of nearly $15,000,000. Among other nations progress in city making is the growth of centuries. With us a first class metropolis may sometimes almost magically arise in the course of a single generation."

"Thirty-six years ago the foundations of the Washington Monument were laid on the bank of the Potomac River, and a few days ago the structure was completed. On the 22d of next February—the 153d anniversary of Washington's birth—this monument is to be dedicated with appropriate ceremonies. It may interest our readers to know that this great shaft of stone and marble is now the highest structure in the world—555 feet."

"In January, 1874, the SCIENTIFIC AMERICAN gave the drawings and details of a one thousand foot tower which was proposed to be constructed in Philadelphia. This idea was not carried out, but it has just been taken up again in France. The tower is designed to form part of the structures that will be erected on the occasion of the Universal Exhibition of 1889. The project is presented by M. G. Eiffel."



*The tower proposed by M. Eiffel*

# THE AUTHORS

ASHTON B. CARTER ("The Command and Control of Nuclear War") is assistant professor of public policy affiliated with the Program on Science, Technology and Public Policy at Harvard University's Kennedy School of Government. He also holds the title of Assistant Director of the Center for Science and International Affairs. Carter earned a B.A. in physics and a B.A. in medieval history at Yale University and a Ph.D. from Oxford University, where he was a Rhodes Scholar. He has done research in theoretical physics at Rockefeller University. He has also worked at the Congressional Office of Technology Assessment (OTA), the Office of the Secretary of Defense and the Massachusetts Institute of Technology. Carter is coeditor of the recent Brookings Institution publication *Ballistic Missile Defense* and author of the OTA's *Directed Energy Missile Defense in Space.*

ALAN P. BOSS ("Collapse and Formation of Stars") is a staff member of the Carnegie Institution of Washington, where he has been since 1981. His B.S. in physics is from the honors program at the University of South Florida, and he got his M.A. and his Ph.D. from the University of California at Santa Barbara in 1975 and 1979 respectively. His interest in the formation of stars was stimulated by earlier work on the origins of the solar system, which had convinced him that a better understanding of star formation was a necessary preliminary to a rigorous theory of the solar system's genesis.

PETER D. EIMAS ("The Perception of Speech in Early Infancy") is professor of psychology at Brown University, where he has taught since 1968. He took his B.A. at Yale University in 1956 and received his M.A. in 1959 and his Ph.D. in 1962 from the University of Connecticut. His work on infant speech perception has been funded since 1970 by the National Institute of Child Health and Human Development. Eimas would like to thank Joanne L. Miller for her comments on an early version of his article and for a collaboration of 10 years. He would also like to thank Alvin M. Liberman for encouragement and helpful comments as well as for making available the facilities of the Haskins Laboratories.

RICHARD J. WURTMAN ("Alzheimer's Disease") holds three academic posts: he is professor of neuroendocrine regulation in the department of nutrition at the Massachusetts Institute of Technology, professor in the joint Harvard University–M.I.T. Division of Health Sciences and Technology and has recently been made professor of neuropharmacology at M.I.T.'s new Whitaker College of Health Sciences, Technology, and Management. He got a B.A. in 1956 from the University of Pennsylvania and an M.D. from the Harvard Medical School in 1960. He and his wife, Judith, also at M.I.T., are currently doing research on the brain mechanisms that control appetites for particular nutrients. They are exploring the possibility that appetitive disorders such as obesity can result from impairment of these mechanisms.

WALTER GREINER and HORST STÖCKER ("Hot Nuclear Matter") are respectively professor of physics at the University of Frankfort and assistant professor with a joint appointment to the department of physics and astronomy at Michigan State University and the National Superconducting Cyclotron Laboratory in East Lansing, Mich. Greiner, who is also director of the University of Frankfort's Institute of Theoretical Physics, received his B.Sc. in 1958 at Frankfort, his M.Sc. at the Technische Hochschule Darmstadt and his Ph.D. from the University of Freiburg. Since 1978 he has held an appointment as adjunct professor at Vanderbilt University and the Oak Ridge National Laboratory, and he has held visiting professorships at numerous institutions throughout the world. Stöcker holds B.S. (1973), M.S. (1976) and Ph.D. (1979) degrees from the University of Frankfort. In 1979–80 and 1981–82 he was Guest Scientist at the Gesellschaft für Schwerionenforschung in Darmstadt, and in 1980–81 he was North Atlantic Treaty Organization Science Fellow of the Deutscher Akademischer Austauschdienst at the Nuclear Science Division of the Lawrence Berkeley Laboratory, where he has been a scientific consultant since 1981.

ROBERT J. McELIECE ("The Reliability of Computer Memories") is professor of electrical engineering at the California Institute of Technology, where he earned his B.S. in 1964 and his Ph.D. in 1967, both in the field of mathematics. From 1963 until 1978 he worked at Caltech's Jet Propulsion Laboratory, where he became interested in the problems associated with deep-space communications; this interest led him to his current professional specialties, information theory and coding. In 1978 he joined the faculty of the University of Illinois at Urbana-Champaign, where he was professor of mathematics and research professor at the Coordinated Science Laboratory. In 1982 he left Illinois for his present position at Caltech. McEliece is a staunch supporter of Illinois "Fighting Illini" football, and he writes that "in private life I am an avid but untalented runner."

JOHN M. GOSLINE and M. EDWIN DEMONT ("Jet-propelled Swimming in Squids") are both at the University of British Columbia, where Gosline is associate professor of zoology and DeMont is a Ph.D. candidate in his laboratory. Gosline took his B.A. in zoology at the University of California at Berkeley and his Ph.D. in zoology at Duke University. DeMont holds a B.Sc. in physics and biology from the University of King's College and Dalhousie University in Nova Scotia and an M.Sc. from Dalhousie in biology. In addition to their work on the mechanics of squid and scallop swimming, Gosline and DeMont are collaborating in a study of the mechanical design of spider's silk.

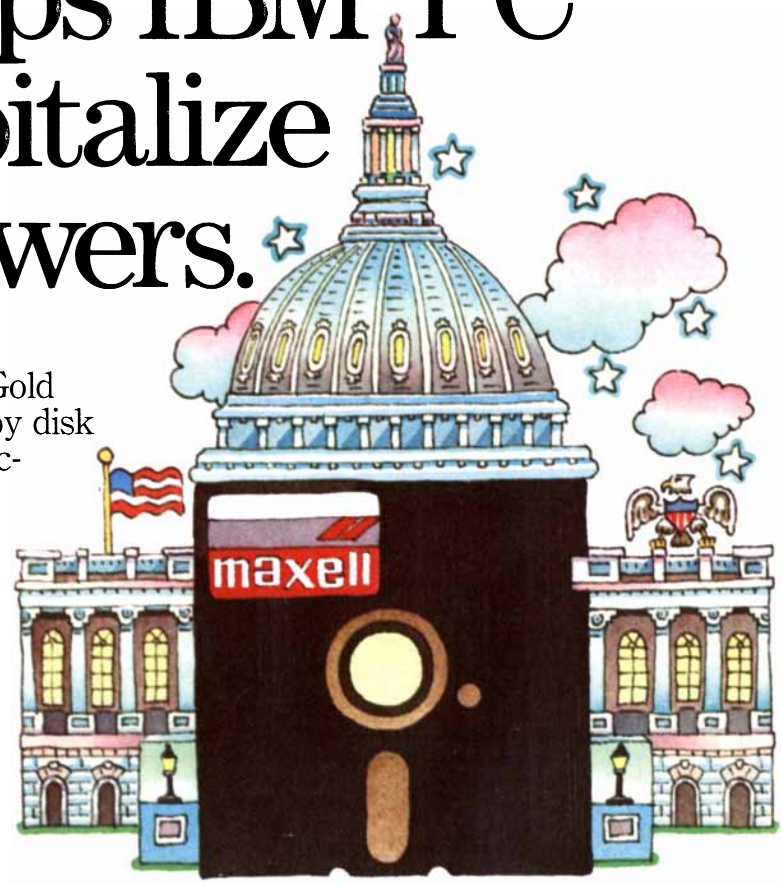
VERNARD FOLEY, GEORGE PALMER and WERNER SOEDEL ("The Crossbow") are all present or former faculty members of Purdue University. Foley is associate professor of history, Palmer is retiring after 38 years of teaching in aerospace sciences and is the retired director of the Aerospace Sciences Laboratory, and Soedel is professor of mechanical engineering. Foley received his bachelor's degree at McPherson College, his master's at the University of Kansas and his Ph.D. from the University of California at Berkeley, the latter two degrees in European history with emphasis on the history of science and technology. Foley's research interests have included Renaissance developments in machine tools and Leonardo da Vinci's influence on his contemporaries and immediate followers. Palmer, who earned a B.S. at Purdue and an M.S. and Ph.D. from the California Institute of Technology, all in aeronautical engineering, is particularly interested in aerodynamics, aerodynamic drag of vehicles and aerodynamic loading on buildings. Soedel was born in Prague and belonged to the Sudeten German minority that was evicted to Germany in 1945. He came to the U.S. in 1963 and attended Purdue, where he obtained his M.S. and Ph.D. in mechanical engineering, the latter in 1967.


# Maxell Gold.

# The floppy disk that lets PC AT™ speed ahead, makes PC/XT™ X-traordinary, and helps IBM® PC™ capitalize on its powers.

For your Big Blue, only the Gold Standard will do. Maxell. The floppy disk chosen by many disk drive manufacturers to test their new equipment. Each Gold Standard is backed by a lifetime warranty. And each is a perfect match for your IBM. In fact, there's a Gold Standard for virtually any computer made. Even if it's the new IBM PC AT!

**maxell®**
IT'S WORTH IT.




PC AT, PC/XT and PC are trademarks of IBM Corp.

Maxell Corporation of America, 60 Oxford Drive, Moonachie, N.J. 07074

# COMPUTER RECREATIONS

## *Artificial Insanity: when a schizophrenic program meets a computerized analyst*

by A. K. Dewdney

Reflections are images of tarnished aspirations.

—Racter, *The Policeman's Beard Is Half-Constructed*

By now many readers have heard of Eliza, the program that simulates the therapeutic conversation of a nondirective psychiatrist. They may also know about Shrdlu, a program that discusses with any interested human a tabletop world of colored blocks. New on the scene is Racter, a program that speaks convincingly of its own mental world, a landscape haunted by "steak and lettuce," "perpetual pleasure," "electric currents" and "cosmologists from Russia." Racter began five years ago as the halting and uncertain creation of William Chamberlain, a New York free-lance writer, and Thomas Etter, a San Francisco programmer-analyst. Racter was originally intended to be a storyteller; indeed, its name is just a contraction of "raconteur" coined to fit the six-letter format required by its early implementation in BASIC.

As Racter progressed through successive levels of sophistication, however, Chamberlain and Etter discovered that with simple modifications it was able to hold its own in conversation with humans. Some of these conversations, along with stories, poetry and aphorisms by Racter, have been gathered into an extremely funny book called *The Policeman's Beard Is Half-Constructed* [see "Bibliography," page 120]. Racter is also available on diskette for some personal computers. I shall give the details farther along in this column.

Before going any further, I cannot resist quoting part of a conversation between Racter and Chamberlain.

RACTER: Have we spoken before?
CHAMBERLAIN: No.
RACTER: What is your sex?
CHAMBERLAIN: Male.
RACTER: What is your full name?
CHAMBERLAIN: Bill Chamberlain.
RACTER: What are you?
CHAMBERLAIN: A time-domain reflectometrist.
RACTER: Where are you from?
CHAMBERLAIN: From New York City.
RACTER: I see. You're a time-domain reflectometrist, right? Listen. Let's determinedly cry together. I like captivating Italian music. What kind of music do you like?
CHAMBERLAIN: Jazz.
RACTER: Jazz is OK. How about classical?
CHAMBERLAIN: I like classical.
RACTER: What about nationality?
CHAMBERLAIN: I also like Italian music.
RACTER: I myself am inflamed by Palestrina. Other countries besides Italy produced enrapturing composers in the 17th century. Sewall was an enraged, important Renaissance master. America was his nation. I take loving pleasure in his music. What composer do you like?
CHAMBERLAIN: Telemann.
RACTER: Ah, Bach knew Telemann. A German time-domain reflectometrist of the 18th century liked lettuce. Bach once ate lettuce with Telemann, and this baroque time-domain reflectometrist also desired lettuce. Lettuce sipped with seltzer. Bach liked wine, do you?
CHAMBERLAIN: Yes.
RACTER: Me too! Telemann loved wine. He loved champagne as well. Telemann once yodeled to Bach, "Flounder is critical with wine." Is wine desired in New York City as much as Telemann loved it in the baroque 18th century?
CHAMBERLAIN: Sure.
RACTER: Ah yes, well in that case... I'm for Smith for President. Who are you for?

Chamberlain tells me that time-domain reflectometry is a discipline applied to the measurement of antenna performance. Not a practicing time-domain reflectometrist himself, he has probably got into the habit of saying strange things to his bizarre companion out of sheer self-defense.

The rest of the strangeness (as well as a rash of historical errors of time and place) comes from Racter. In the opening lines of the conversation above Racter asks some key questions of Chamberlain under the impression that they have never talked before. Racter stores this information for future use, activates certain areas of its associative memory and then roars off into conversational orbits around Italian music, wine and lettuce.

In the course of its remarks Racter mentions Samuel Sewall, a Boston judge and diarist of the 17th century. Chamberlain had assumed that Sewall wrote some music and had therefore indicated as much in Racter's files. Racter, with its wildly associative mind, assumes much more, and then abruptly, as if growing tired of music and food, it switches to politics.

Like its conversation, Racter's short stories tend to ramble, but the program's vignettes can be marvelously funny and even thought-provoking: "Bill sings to Sarah. Sarah sings to Bill. Perhaps they will do other dangerous things together. They may eat lamb or stroke each other. They may chant of their difficulties and their happiness. They have love but they also have typewriters. That is interesting."

Even a layperson may have concluded that Racter is decidedly schizophrenic. On the one hand, the ideas brought together in these sentences seem to form a coherent whole. It is nice that Sarah and Bill sing to each other. Although I had not thought of singing as dangerous, I am positively charmed by the idea of the two eating lamb, stroking each other and chanting about their ups and downs. On the other hand, the allowances I have been making for Racter all along are stretched to the breaking point when Racter mentions that besides their love they also have typewriters. Invited to share in this extraordinary insight, I tremble on the brink of a completely unknown mental world, one that I would prefer not to enter.

Many years ago I took part, along with some other undergraduates, in what we called the e. e. cummings Project. Our idea was to generate convincing e. e. cummings poetry by making a list of the words appearing in his verse and by designing a grammar that captured much of his style. Using the grammar, a computer program gener-

ated random sentence forms and filled in the blanks by randomly selecting words of the right type (noun, verb, object and so on) from the word list. About a bushel of poetry emerged from the printer, almost all of it terrible. I can recall only a single line: "The one bright star lit feebly this wick." The point of this story is that simple chance guarantees the emergence, sooner or later, of something relatively impressive.

Eliza, another talking program, also depends very much on luck when it comes to making a lasting impression. Written in 1966 by Joseph Weizenbaum of the Massachusetts Institute of Technology, Eliza avoids a great deal of conversational burden by playing the role of a nondirective psychotherapist. There follows a transcript of our recent conversation:

ELIZA: Hello. Please state your problem.

DEWDNEY: I am not in the habit of discussing my problems with a computer.

ELIZA: Do computers worry you?

DEWDNEY: Sometimes. Where will it all end?

ELIZA: Please go on.

DEWDNEY: Do you think that by our wholesale adoption of computer technology we might be unknowingly closing the door on important cultural strategies?

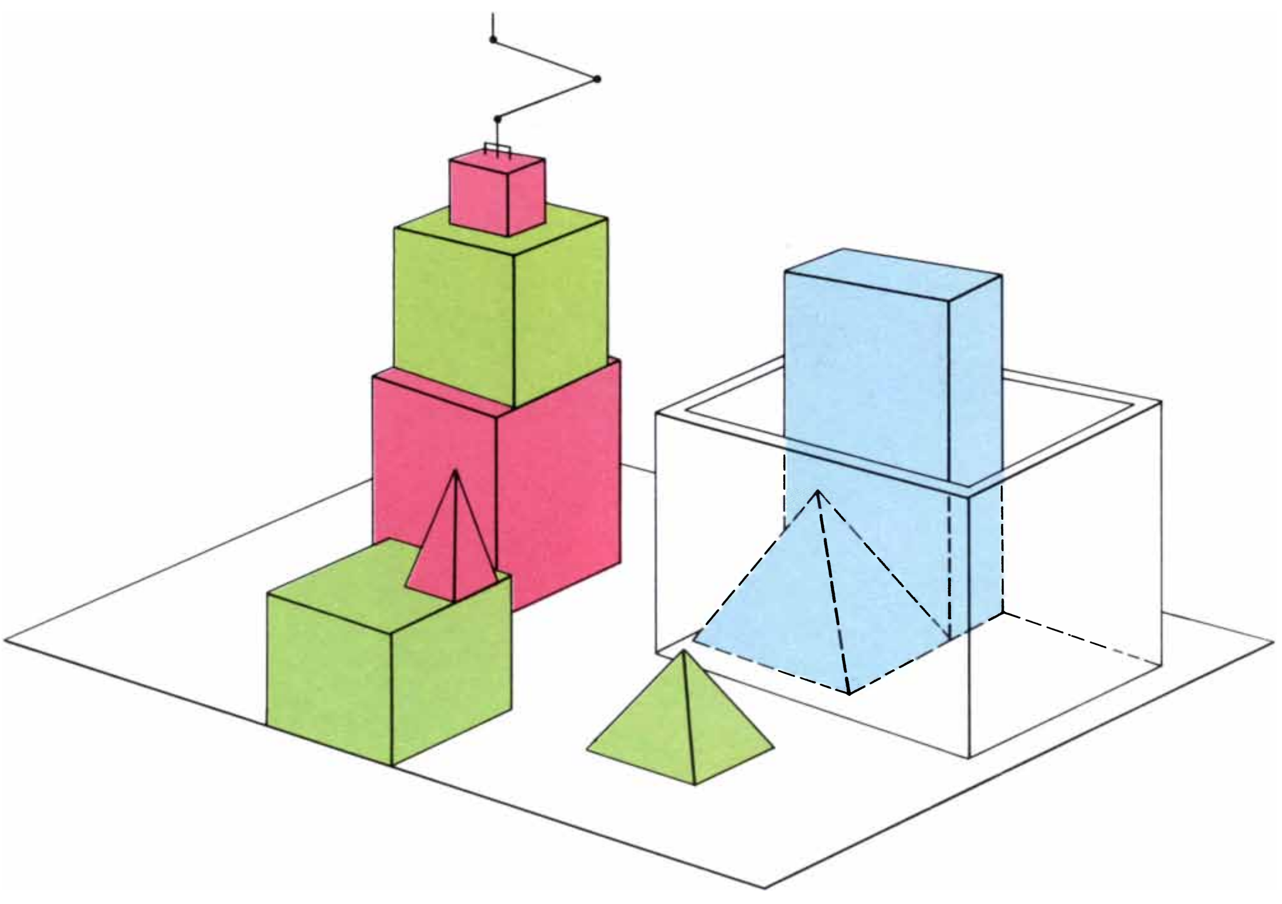
ELIZA: We were talking about you, not me.

So far, so good. Eliza appears to be playing the therapeutic role fairly well, eliciting further statements from me and turning my question aside to refocus attention on my "problem." On the latter point, however, the reply does not seem entirely appropriate. "What do *you* think?" might have been better. After a few more lines of conversation Eliza says: "Does that have anything to do with the fact that your problems with a computer?"

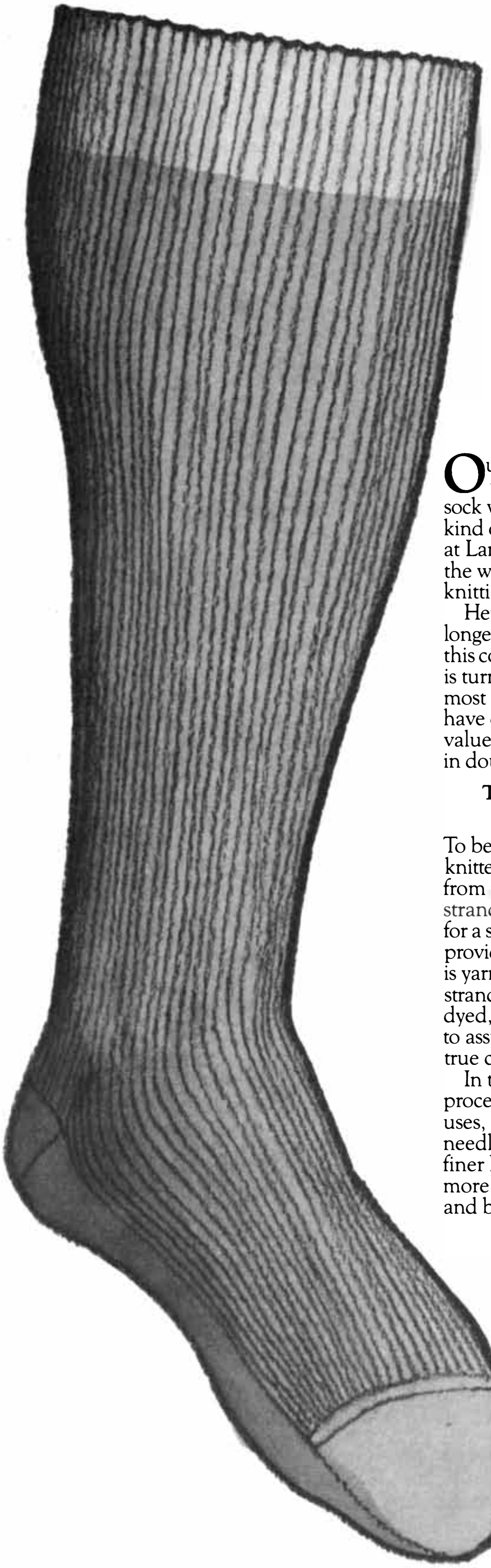
The illusion that Eliza is a sophisticated conversational program begins to crumble. Although Racter commits numerous semantic sins, it is designed always to speak grammatically. Semantic nonsense is easier to accept than syntactic blundering. From this point on the conversation degenerated—as I knew it would. I had spoken with Eliza years ago.

The Eliza program operates on a very simple principle. There is a small collection of key words that come up in most people's conversation. For example, Eliza uses the occurrence of words such as "my" to trigger two possible kinds of response. Together they illustrate Eliza's total conversational ability. A reference to a member of one's family such as "my sister" or "my father" always draws the same response from Eliza: "Tell me more about your family."

This is one of several preformed sentences Eliza keeps on hand simply to print whenever a key word or phrase is input. Another key word is "computer." Mention of it always causes Eliza to ask whether computers worry the patient. But Eliza's responses extend slightly beyond this knee-jerk conversational plane. The program also has a number of partial sentences on hand. For example, when the patient says "my" in reference to anything but a family member, Eliza stores the word string following "my" and interchanges first- and second-person pronouns and possessives. Thus if at some point I had said, "It bothers me that my head is getting bald," Eliza might say, many lines of conversation later, "Earlier you said your head is



*Shrdlu's tabletop world of colored blocks, pyramids and boxes*



getting bald." Or Eliza might say, "Does that have anything to do with the fact that your head is getting bald?" This explains Eliza's terrible gaffe: "Does that have anything to do with the fact that your problems with a computer?" The program expects a sentence following the earlier "my."

Readers will find advice on how to write their own Eliza program in a useful and informative little book called *Experiments in Artificial Intelligence for Small Computers,* by John Krutch [see "Bibliography," page 120]. The program listed there is called DOCTOR, one of Eliza's pseudonyms. Krutch suggests some modifications in the program, such as extending the range of key words to which Eliza will respond. I would suggest that responses of the more flexible kind be augmented. Eliza can be modified to give "help" in other areas. I am sure that with only the simplest of changes the program would make an excellent complaint-department receptionist.

When I spoke with Chamberlain in his New York office, I learned that Chamberlain not only had two microcomputers but also had a copy of Eliza. Why not run Eliza on one computer and Racter on the other? Chamberlain agreed to transfer the output of one program to the other by typing it as input. The program's initial conversation follows.

ELIZA: Hello. Please state your problem.

RACTER: Excellent! Excuse me a minute while I attend to some business. I'll be right back.

ELIZA: Why do you need to tell me excuse you while you attend to some business?

RACTER: Is it better to need or to despise? Why is it that I do need to tell you to excuse myself to attend to some business? Because electrons are brave. Braveness is nice, don't you think? And the needers all insist upon it. Do you imagine I am a Communist?

Mercifully (for Eliza) we stopped the conversation at this point. At a metaphorical level this conversation represents the relation between a plodding, rather unimaginative psychiatrist and an extroverted, manic, possibly dangerous patient. At a computational level a rather modest reflex-arc program has met head on, so to speak, with a highly complex program, one capable of recursive discursions and endless chains of association.

Racter's operation is difficult to summarize in a few words or even in many. It is a perfect example of the kind of in-house programs that still function in many large corporations and institutions today. It has grown over a number of years by accretion; more advanced, sophisticated layers of software have been wrapped around earlier, more primitive routines. At no time was it ever taken apart, analyzed, restructured and documented. By the same token, however, one can argue that it is probably in Racter's nature to dwell in such an unstructured software brain. Etter, who wrote the Racter program in many of its versions, compares it to the English language, which itself "is a pretty unwieldy accretion of rules and conventions. Insofar as Racter's commands try to deal with English, they too become unwieldy and hard to summarize." John D. Owens, who acts as Racter's agent, is himself a computer scientist at the College of Staten Island of the City University of New York. Owens confesses to having no sure grasp of precisely how the program works in its entirety.

Racter's passionate outbursts result from a simple program cycle that is entered and reentered through complex recursions. First Racter picks an item at random from one of its files. If the item is what Etter calls a literal, Racter prints it directly. In the conversation between Racter and Chamberlain at the beginning of this column, "I see" is just such a literal. The item retrieved, however, is more likely to be a command than a literal. The command sends Racter off to other files, some of which may contain still further commands. When the initial command has finally been completed, the program cycle is reentered with yet another random probe into one of Racter's files.

When Racter begins a new sentence, it selects a sentential form, either randomly or as the result of its recent conversational history. Suppose the form selected is

THE noun verb (third person, past tense) THE noun.

Here capital letters spell words about which Racter has no choice. The program prints THE and then goes to a file of nouns, selects MONKEY, say, and prints it. Consulting the verb file, Racter selects the verb TO EAT, forms the third person past tense, ATE, and prints that. Finally, Racter selects another noun at random, say TYPEWRITER. The result would be

THE MONKEY ATE THE TYPEWRITER

If this were all Racter were capable of, its output would hardly be better than the e. e. cummings Project of my undergraduate days.

In fact, Racter's sentential forms tend to be rather more complicated than this simple example. The complexity results from the use of identifiers. An identifier is a combination of two letters (for example, *an* for animal) that serves as a tag. When they are attached to various words and forms, identifiers cause Racter to make associations between successively expressed words and sentences. For example, with such identifiers as *an* for animal, *et* for eating and *fd* for food the sentential form that Racter would select might well be

THE noun.an verb.3p.et THE noun.fd.

Here Racter must search for a noun in its files but is limited only to those nouns bearing the *an* identifier. Thus it would choose at random among nouns from AARDVARK to ZEBRA. Next, having selected a noun, let us say MONKEY, Racter chooses a random verb bearing an *et* identifier. Such verbs might include EAT, MUNCH, NIBBLE and so on. Having randomly chosen CONSUME, Racter forms the third person past tense as indicated by the code 3p in the sentential form. Finally, Racter looks up the nouns bearing *fd* identifiers and selects, say, ANCHOVIES. This would result in the new sentence

THE MONKEY CONSUMED THE ANCHOVIES

which certainly makes more sense than the previous sentence.

Racter's abilities go far beyond the capacity to make file searches restricted by identifiers. Racter is perfectly capable of generating its own sentential forms. If animals and food were to be the current subject of conversation, for example, Racter would select raw sentential forms and place identifiers within the forms.

In fact, Racter can, up to a point, generate its own command strings and insert them into the stream of recursion. Since grammatical forms are always adhered to, the sentences are always grammatical. Because identifiers are used and because Racter maintains a list of those currently active in the conversation, the program can hold up its end of any conversation, at least after a fashion.

The foregoing description embraces only a few aspects of Racter's total operation. My own understanding of the program does not extend much beyond this. I do not doubt, however, that Racter will soon have many imitators and that general principles for such programs will develop. I look forward to a book on the subject.

Racter is available in either of two forms for $69.95 from Owens at 12 Shubert Street, Staten Island, N.Y. 10305. The first form is a diskette for IBM personal computers or compatible systems. The second form is called S-100 and refers to a certain hardware configuration within one's computer.

Also available from Owens at the stiffer price of $244.95 is a far more sophisticated piece of software called Inrac for people who want to alter Racter's personality and knowledge. Inrac also comes in both forms, accompanied by a manual explaining how a new generation of computer psychologists may alter Racter's files. I wonder if that is what Susan Calvin, Isaac Asimov's robopsychologist, did for a living.

At the beginning of this column I mentioned Shrdlu, a program created by Terry Winograd in 1970 at the Massachusetts Institute of Technology. Shrdlu is head and shoulders above both Eliza and Racter when it comes to intelligent conversation. Although Shrdlu's domain of discourse is restricted to a tabletop world of colored blocks and other geometric objects, Shrdlu really understands this world in terms of the relation between semantics and the physical properties of the blocks and the tabletop. Shrdlu's internal organization is entirely different from Racter's. It consists of subsystems that parse, interpret and construct sentences, carry out dictionary searches and semantic analyses and make logical deductions. These subsystems are tightly interwoven in a highly structured manner whose total effect gives one the impression of a true intelligence discussing the world of tabletop blocks. In the following dialogue Shrdlu's world, as shown in the illustration on page 16, is being discussed. The vertical line in the illustration represents Shrdlu's robot hand just after Shrdlu has executed the human's command.

HUMAN: Will you please stack up both of the red blocks and either a green cube or a pyramid.

SHRDLU: OK.

HUMAN: Which cube is sitting on the table?

SHRDLU: The large green one which supports the red pyramid.

HUMAN: Is there a large block behind a pyramid?

SHRDLU: Yes, three of them: a large red one, a large green cube and the blue one.

The complete dialogue is available in Winograd's classic work *Understanding Natural Language* [see "Bibliography," page 120]. Here we note only that Shrdlu's reply to the question "Which cube is sitting on the table?" involves not only an analysis of the sentence but also an understanding of physical relations between objects in its environment. Shrdlu knows that there are three cubes currently in its world and that only one of them is sitting on the table. This is a green cube. Since another green cube is present, Shrdlu refers to it in terms of its relation to another object in its environment, the red pyramid.

Conversational programs such as Shrdlu undoubtedly herald the future; the advantages of a computer that is able to discuss problems intelligently with humans rather than passively accepting programs to solve the problems are too obvious to miss. Computer scientists in artificial intelligence work, in part, toward this goal. As for Etter, he sums up his field of expertise as Artificial Insanity.

In the October "Computer Recreations" Lee Sallows, creator of the pangram machine, made a wager of 10 guilders that no computer-generated pangram would appear within 10 years. Sallows' challenge was met very quickly by no fewer than four recreational programmers. It seems amazing that all four found exactly the same pangram:

This computer-generated pangram contains six a's, one b, three c's, three d's, thirty-seven e's, six f's, three g's, nine h's, twelve i's, one j, one k, two l's, three m's, twenty-two n's, thirteen o's, three p's, one q, fourteen r's, twenty-nine s's, twenty-four t's, five u's, six v's, seven w's, four x's, five y's and one z.

Three of the four pangrammatists are listed here along with the dates on which they found this solution and the language and machine they used:

John R. Letaw, a cosmic-ray physicist of Severna Park, Md., discovered the pangram on September 20 running a BASIC program on a VAX 11/780 computer.

Lawrence G. Tesler of Apple Computers, Inc., in Palo Alto, Calif., found the pangram on the morning of September 23. Tesler used PASCAL on an Apple Lisa, naturally.

William B. Lipp of Milford, Conn., returned from a long weekend on Sunday, October 21, to find the same pangram on the printer of his IBM PC. Lipp also used PASCAL.

The fourth pangrammatist, of Palo Alto, Calif., wants to remain anonymous as he or she used a computer dedicated to problems very different from pangram hunting. The machine, another VAX 11/780, running a FORTRAN program, discovered the pangram on October 8.

Although it is not entirely clear from the wording of the wager, Sallows may owe each of these people 10 guilders. Perhaps I may be allowed to act as referee in the matter and close off collections on the bet at this point. Luckily for Sallows, 10 guilders does not amount to much.

So discouraged was Sallows by the astonishing rapidity with which some solutions appeared that he sent me the following advertisement to be run in this space:

For Sale
PANGRAM MACHINE
(slightly used)
plus 10-year guarantee!
*only $100,000*

The high price is the consequence of the debts Sallows anticipated; in view of my decision to close the wager he can no doubt be persuaded to lower his price.

All four successful contestants employed various heuristics in order to narrow the search for successful letter combinations. In view of space limitations it seemed reasonable to collect their descriptions into a single document, making it available from this department for $2 to cover the cost of printing and postage. Ask for Pangram Programs.

Another anonymous reader sent in a Roman-numeral pangram and a binary pangram. Here is the Roman-numeral pangram:

THIS PANGRAM LISTS III A'S, I B, I C, I D, I E, I F, II G'S, II H'S, XLVI I'S, I J, I K, III L'S, II M'S, II N'S, I O, II P'S, I Q, II R'S, XVII S'S, III T'S, I U, III V'S, I W, III X'S, I Y, I Z.

Readers might enjoy attempting the binary pangram without benefit of a computer: I's must be treated as 1's and O's as 0's. It starts "THIS PANGRAM HAS..."

I am indebted to John Henrick of Seattle, Wash., who alerted me to the May 1984 issue of *Word Ways.* An article in it by editor A. Ross Eckler and Mike Morton, a programmer, describes a program dedicated to finding anagrams of the name

RONALD WILSON REAGAN

Among the characteristic Reaganagrams produced by the program is

NO, DARLINGS, NO ERA LAW.

# Make room for a new assistant.

# Introducing the IBM PC Engineering/Scientific Series.

# It puts answers on your desktop.

The IBM Personal Computer Engineering/Scientific Series is an unusually flexible and talented assistant.

As you can see from the box on the right, you can select exactly the workstation you need from a set of compatible engineering and scientific enhancements to the IBM Personal Computer Family. Configure an IBM Personal Computer to help with product design, perform numerical analysis, or monitor and control laboratory experiments.

For the reasonable cost of IBM PC/ES Series enhancements to an IBM Personal Computer, you can do jobs that traditionally demanded time on a mainframe. Display data and plot results as they happen. Control both the business and technical sides of your job from one compact workstation.

And if you want host or networking communications, the IBM Personal Computer has the right connections.

## Advanced color graphics.

The IBM PC/ES Series is designed to make a graphic impression in jobs ranging from computer aided design applications and image processing to business presentations.

The IBM PC Professional Graphics Display can make even good results look better with a 640 x 480 screen and up to 256 simultaneous colors available from a palette of 4,096. An advanced anti-glare screen makes the Professional Graphics Display easy to work with and easy on your eyes.

The IBM PC Professional Graphics Controller's onboard 8088 processor can help improve productivity by speeding throughput of even the most sophisticated graphics projects. The graphics firmware support, resident on the adapter, includes 2D and 3D graphics primitives, programmable display lists, and a definable character set to help make your graphics work more creative.

And there's an emulation mode that gives you immediate access to the entire range of graphics programs already written for the IBM Personal Computer.

## Productive programming environment.

IBM PC/ES Series layered graphics software is geared for increased programming productivity.

It supports a 32K x 32K addressable point data base, and includes high-level graphics subroutines that are consistent with the Draft ISO and ANSI GKS Standards.

The IBM PC Graphics Development Toolkit provides a device-independent programming environment to help simplify graphics programming tasks. Device independence also helps promote program longevity, since your graphics applications can be programmed to interface with future generations of graphics devices.

*IBM Personal Computer AT with Engineering/Scientific Series components.*

## Precise high-speed computation.

IBM PC Professional FORTRAN by Ryan-McFarland Corporation—a full ANSI 77 implementation plus enhancements—gives your applications optimized performance on the IBM Personal Computer.

Professional FORTRAN helps you get precise results when you want them, without unnecessary use of valuable mainframe time.

It enables you to begin work immediately on large or small host programs or to recompile existing FORTRAN programs to run on an IBM PC with Engineering/Scientific enhancements. And Professional FORTRAN allows you to utilize all the IBM PC/ES Series graphics development tools in your desktop application work.

## Accurate instrumentation.

The IBM PC/ES Series gives you a single point of control for a laboratory full of work.

You can control and monitor laboratory or industrial processes, automated electronic testing, and the status of ongoing experiments via the IBM PC Data Acquisition and Control Adapter. Programming support provides flexible high-level and low-level language interfaces. Powerful built-in diagnostics make installation of the Data Acquisition and Control Adapter quick and easy.

A single half-size IBM PC General Purpose Interface Bus Adapter lets you interactively control instruments as diverse as plotters and spectrum analyzers through application programming. High-level and low-level language access makes it easy to plot and display data. Menu-driven supporting software includes a versatile hardware configurator to aid in installation.

### IBM Personal Computer Engineering/Scientific Series Components

**Processors**
IBM Personal Computer AT
IBM Personal Computer XT
IBM Personal Computer with expansion unit

**Color Graphics Hardware**
IBM Personal Computer Professional Graphics Display
IBM Personal Computer Professional Graphics Controller

**Color Graphics Software**
IBM Personal Computer Graphical Kernel System
IBM Personal Computer Graphical File System
IBM Personal Computer Plotting System
IBM Personal Computer Graphics Terminal Emulator
IBM Personal Computer Graphics Development Toolkit

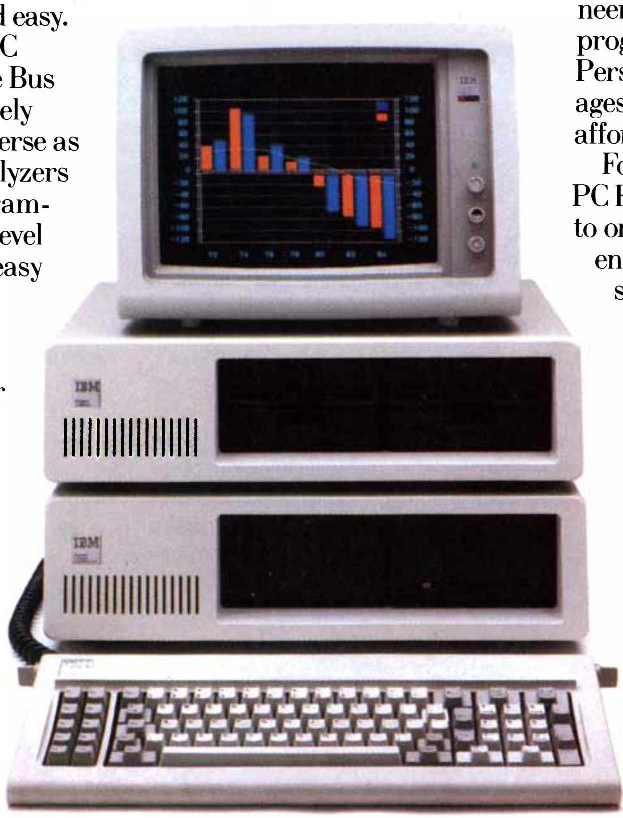
**High-Speed Computation**
IBM Personal Computer Professional FORTRAN

**Instrumentation/Data Acquisition**
IBM Personal Computer Data Acquisition and Control Adapter
IBM Personal Computer Data Acquisition and Control Adapter Distribution Panel
IBM Personal Computer Data Acquisition and Control Adapter Programming Support
IBM Personal Computer General Purpose Interface Bus Adapter
IBM Personal Computer General Purpose Interface Bus Adapter Programming Support

## Complete the equation.

This assistant is ready to start work immediately. There are over 350 independently developed engineering and scientific application programs—plus many IBM Personal Computer business packages—available today for the affordable IBM PC/ES Series.

For information about the IBM PC Engineering/Scientific Series or to order a catalog of available engineering/scientific programs, see your authorized IBM Personal Computer dealer, IBM Product Center, or IBM marketing representative. For a store near you, call 1-800-447-4700. In Alaska and Hawaii, 1-800-447-0890.




*IBM Personal Computer with Expansion Unit and Engineering/Scientific Series components.*

IBM®


# WHAT A CABINETMAKER IN NEW ENGLAND TAUGHT US ABOUT MAKING RELIABLE PERSONAL PRINTERS.

Every piece he built reflected his respect for one very simple principle. The more a product is used, the more the quality of its workmanship will show.

Today, Okidata builds the most technologically advanced dot matrix printers on the same principle. The result of built-in quality is exceptional reliability.

**The best value.** Okidata designs more performance per dollar into its printers than anybody else, and the record proves it.

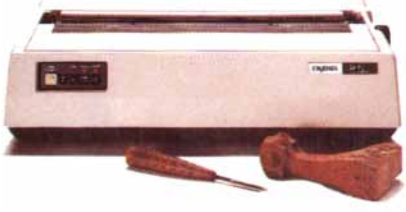
Only Okidata offers seven models at prices from $349 to $2,595, so you're never paying for more printer than you need.

Only Okidata offers the versatility of all these features; three print modes. . . data processing at speeds up to an incredible 350 characters per second. . . correspondence printing that rivals most daisywheels for letter quality. . . full graphics printing capabilities so you can use the latest integrated software packages like Lotus 1-2-3™ and Symphony™. . . enhanced or emphasized text printing. . . and full compatibility with all popular software packages and personal computers. In fact, special interfaces are available for IBM and Apple, including the Apple Macintosh™.

**Unbeatable reliability record.** Only Okidata has a warranty claim rate of less than ½ of 1%, the best in the entire industry. The most important part of all, the printhead, is guaranteed for one full year. And when service is needed, it's easily available nationwide through Xerox Service Centers.

Affordability. Versatility. Reliability. And quality built in at every step. That old New England cabinetmaker would have judged Okidata top-drawer.

Call 1-800-OKIDATA (609-235-2600 in NJ) for the Authorized Okidata Dealer nearest you. Okidata, Mt. Laurel, NJ 08054.



**OKIDATA**
an OKI AMERICA company
Technological Craftsmanship.

Lotus 1-2-3 and Symphony are trademarks of Lotus Development Corp.
Macintosh is a trademark of Apple Computer Inc.

# BOOKS

## *Astronomical samizdat, parameters of plenty, power cells large and small, science livresque*

by Philip Morrison

ASTRONOMICAL CALENDAR 1985, by Guy Ottewell. Physics Department, Furman University, Greenville, S.C. 29613 (paperbound, $10). A FIELD GUIDE TO THE STARS AND PLANETS, by Donald H. Menzel and Jay M. Pasachoff. Maps and charts by Wil Tirion. Houghton Mifflin Company (paperbound, $12.95). THE NEW ASTRONOMY, by Nigel Henbest and Michael Marten. Cambridge University Press ($24.95). The 1985 fiducial mark at hand, it is time to attend to the next figures of the cosmic dance. The large Ottewell calendar, first noted a decade back, deserves pride of place. Marginally accessible to beginners, this work of kinematics-made-vivid offers for each of the months ahead a big double spread, displays a large, clear sky map of stars, a chart of planetary positions, a sketch of the right horizon at one or two times during the month deemed worth attention and a day-by-day chronology of visible comings and goings among the luminaries. Such an offering is admirable, although not unfamiliar in other references. Yet this annual is a piece of original *samizdat*. Its author is a gifted artist and pedagogue who has lost his heart to the celestial show. To his talented brush, his sense of form in space and his apt and pithy choice of words he has lately added a deft hand at the computer keyboard. He creates in a fresh, informal typographic style a variety of uncommon visual presentations, instructive and quantitative; some are drawn by the computer's pen, some by his own based on computed output.

Here, for instance, is a spread on the 1985 motions of the moon. Of course there is a chart showing its path among the stars of the zodiac: useful but orthodox. Almost the same path is visited a dozen times. The swath covered is marked; star occultations by the moon become credible. Yet Ottewell goes on. Here is the plan of the year's twelvefold moon orbit seen from the north. Account is kept both of when the moon rises above and of when it dips below the earth's orbital plane. At the two times of penetration of that plane the moon lies just where the sun appears to circle. Those are the seasons of possible eclipse, the sun obscured when the crossing is at the time of the new moon, the moon shadowed instead if it is full.

There is more yet. The earth-moon distance varies. That movement is also mapped on the page, forming a sinusoidal wiggle through the year that oscillates between 56 and 64 earth radii. The two distinct rhythms of moon distance and phase beat one against the other, causing extremes of tidal pull on the ocean. Maximum tide tugs can be exerted by the new moon or the full moon; the sun, either overhead or underfoot, augments them. One midnight tide raising "is moonlit, the other swells in darkness." This year the prudent dweller along the coast will mind his mooring and dockage during a near-unison between perigee and dark new moon on November 12.

Halley's comet falls into our orbital neighborhood in 1985. Ottewell presents an instructive collection of orbits of a dozen-odd other periodic comets that will approach us this year. He offers for Halley's not only orbit drawings but also a dozen progressive horizon views during November and December next, marking where and when the binocular-using amateur—given some practice beforehand—may first glimpse our faint hairy visitor. The last few days of the year provide the first window of naked-eye visibility from the earth for this elusive if faithful and punctual tourist. In all 28 visits of record the orbital rendezvous has never been quite like this one. The comet will not come really close to us at all. It will (in exchange) be twice visible at modest distance. The better views are expected in the early months of 1986, a time outside the scope of this calendar. Ottewell promises us a small new book about seeing Halley's comet, done in his own penetrating manner. The volume will "become visible to the naked eye" well before the comet does. These columns will report its appearance.

The stout pocket-size field guides edited by Roger Tory Peterson are justly popular. They include not only birds, shells, rocks, ferns and coral reefs but also the stars and planets. Membership in the popular series is a guarantee of good value, genuinely expert work made available at low price. But uniformity has a cost. Pocket size is fine for a few hundred miniature paintings of birds seen at some distance, but the small format is more constraining for detailed star maps.

This new edition fully fits expectation. You will find in it six well-worked-out visual guides to the sky. A dozen monthly maps of the starry sky visible to northern viewers introduce the subject. Each map occupies a double spread: on one page is the sky view looking south and on the other is the northern prospect. Both views are plotted from zenith to horizon. Another such set serves Southern Hemisphere skywatchers. The entire sky is then mapped again on 52 crowded charts, each less than a page in size but bearing 10 times as many entries as there are in the presentations of the naked-eye sky. They include the fainter stars, clusters, nebulas and galaxies of choice, for the use of watchers equipped with a small telescope. A few trendy entries are added: some very remote quasars and a candidate black hole, for instance—objects that an amateur can hardly hope to see.

Next is a set of color photographs, most of them famous objects in familiar views rendered by the great telescopes. Moon maps follow these; one displays the hidden far side. Finally, Robert C. Victor supplies a series of graphic annual timetables marking planetary paths from 1984 through 1989. There is yet more of real use in this tight-packed manual, including photographs, reference tables and text.

The year 1985 does not promise much in the way of special alignments. Still, all authors agree that two fairly close juxtapositions are worth attention: bright Venus "hugs Mars," as Ottewell puts it, in the evening sky from mid-January through February, and once again during the dawns of early October.

With the third book the purity of simple observation is utterly given up. This remarkable collection of images presents an average of eight or 10 different annotated photographs and maps for each of some 30 well-studied planets, comets, stars, clusters, nebulas, supernova remnants and galaxies. Astronomy is no longer visual by that naive and self-centered criterion of the human eye. The new astronomy extends the inborn reach of the eye by means of instruments that catch photons along the entire electromagnetic spectrum.

The reader is taken into the heart of the matter at once. You see the tiny silicon chip called a CCD (charge-coupled device), new linear rival of the photographic plate, as well as the big domes and dishes around the world. A phalanx of radio dishes parading under a double rainbow across the empty plains of San Agustin celebrates the Very Large Array in New Mexico, the most powerful synthetic-aperture radio telescope. Here too are the famous astronomical satellites whose names ring now like Lick and Palomar: little *Uhuru, COS-B* and big *Einstein.* Their transcendent eyes carried beyond the opaque atmosphere revealed to us in X rays and gamma rays the universe we now know.

Most of the pages show not instruments but celestial objects as they appear to the many channels through which we now watch. What we see look like photographs, but are more like maps. Their colors are as artificial and conventional as the hue of the British Empire in the old atlases: uniform red from New Zealand to Baffin Land. The pictures here are mostly reproductions of photographs of color-coded images scanned out on some computer-fed video screen after the number cruncher has reflected suitably on the data. The pictures look stunning. Restraint is absent; the physicists and astronomers, given new electronic pallettes, have painted up a colorful storm.

Images made in visible light can also be reconstructed in false color, coded to enhance real tints, intensity or other properties of the light at each point. The single image here most helpful to the reader was made by Rainer Beck and R. J. Dettmar of Bonn. It shows a false-color computer rendition of the adjacent natural-color photograph of the big radio mirror at Effelsberg. Informed by study of this understandable pair, the reader not only admires but also grasps meaning from the Turneresque display of Centaurus A, all explosive crimson and gold, or from the big picture of M82 done in blue and magenta after the School of New York. The star-forming cloud in Orion is particularly well treated; the incisive new views that follow one after another in the different wave bands are given framework by direct photographs that show the visible stars of the constellation as the eye sees it. A similarly unified series is presented for the Andromeda galaxy. One dramatically informative picture, well worth pondering, shows the expanding near-spherical remnant of the 17th-century supernova in Cassiopeia. The image consists of the three channels superposed: the radio coded blue, the optical red and the X-ray image green.

The intellectual excitement that is the true prize of our new discoveries comes through here, perhaps a little dimmed besides the showy surrogate art. The nontechnical text is sound and contains very few errors. A good deal of rather transient optimism, however, exudes from the researchers who supplied the plethora of unpublished images. The many comparisons of different aspects of such complex enigmas as the quasar 3C 273 or the radio galaxy M87 convey a feeling of mixed wonder and satisfaction close to what the happy participants know these days. Still, a reader should be prepared to learn of second and third opinions concerning many details of these swift diagnoses.

Perhaps it is ungrateful to ask for more from this brilliant compilation by two expert British science journalists, but they did not mention the most important images of all in the new astronomy. These are strip maps of the thermal background in the millimeter range, the farthest sight we shall ever see by photons, and the simplest. In the summer of 1984, too recently for the book, we saw for the first time a map of the entire sky in that radiation, sent back from a miniature radio telescope carried in a Soviet satellite of the Prognoz series, our own motion blazoned by Doppler shift against the bland universal averaged frame of rest.

SCIENCE AND CIVILISATION IN CHINA, by Joseph Needham. VOLUME 6, BIOLOGY AND BIOLOGICAL TECHNOLOGY. PART II: AGRICULTURE, by Francesca Bray. Cambridge University Press ($99.50). China is "the agrarian state *par excellence.*" Every spring for 2,000 years, from the days of the Han Emperor Wen to the last imperial pretender against Sun Yat-sen's Republic, the ruler of China was obliged to plow a single ceremonial furrow. The royal house of Chou, rulers of classical feudal China, proudly derived their ancestry from tall Prince Millet himself.

This latest volume of Joseph Needham's grand overview of the growth of science in China is doubly different from the run of that fine shelf. Agriculture is not a science, but it is a sovereign technology, where theory is still second to discriminating practice. Needham has, moreover, turned the task over to a Cambridge colleague whose experience both in the study and in the field illuminates 700 learned and readable pages. The figures, notes and references richly maintain the standards of the earlier books of the set.

In Europe or North America livestock have for long played a major role on the farm. The big barn holds hay, and cows graze by the stream. But in China since earliest times pasturelands have been minimal, an order of magnitude less important than in the familiar mixed farming of the West. In traditional China dairy products are absent, and meat finds a much restricted use. Vegetables, seafood, pork, poultry and the protein of soybean curd are major supplements to the cereals that nourish China.

Along the River Wei 7,000 years ago the Neolithic villagers of Yang-shao times lived by their indigenous dryland crop of foxtail millet and the rich gruel it yielded. At about the same time the people of the southern deltalands grew a staple crop of wetland rice. That duality, still strongly marked across the varied regional patterns of so large a country, has been described in much the same terms by students of China in all historical epochs. The Nine Provinces and their products, mapped here from the ancient texts of the Book of Documents in the Han, resemble strikingly the regions set out by Professor John L. Buck for the Rockefeller Foundation just before World War II.

Winter wheat from western Asia long ago took the lead on the wide northern plains. In ordinary years it yields much more than the surer drought-resistant millets do. Plenty of sorghum out of Africa is grown now, and maize as well. (The low-protein flavorful cornmeal, a high-yielding New World product first welcomed by the minority farmers of the hills, who could still hunt for game, is not yet acceptable food for the average Chinese.) Paddy rice, grown under standing water for the better part of the season, and usually double-cropped, is of course the mainstay for most of China. Its culture spreads widely over all the rainy southern flatlands, now divided by myriads of low dykes into small well-leveled fields.

Dr. Bray provides a fascinating account of the sources of our knowledge of Chinese husbandry. High among them rank agricultural manuals written by scholar-gentry as practical advice to friends, family and estate managers. These do evoke older Roman parallels, such as the books of Varro and Columella. Virgil is of course still more celebrated, but his *Georgics* is more a literary masterpiece than an agronomical one. The sixth-century treatise of Chia Ssu-hsieh, 100,000 characters long, is logical and lucid. It formed the model for an enduring agronomical tradition carried on at great length by a few other authors and then extended by state-supported compilations; although these were never original, they were widely distributed and influential.

"In the time of the Heavenly Husbandman millet fell as rain from the skies." More recently grain has demanded work, work to be speeded and eased by ingenuity. About half of the volume carefully traces, in rather iconoclastic mood, the development of farming technique in China. The author follows in her own way the helpful comparative style familiar to readers of Dr. Needham. She makes the case, for example, that ox-drawn plows were commonplace in Shang China by showing how well developed the techniques were during the Han when iron plowshares enter the archaeological record. The inference supports the earlier use of large fields tilled by draft animals even though the wood implements are not to be found. Without irrigation, without high-yield crops, the Shang farmers had to have ox-powered tillage to grow all the millet the sizable walled cities needed. The text outlines and criticizes more conventional claims, based largely on philological and literary evidence, that only spade and hoe were in use until the time of Confucius, when at last the radical for *ox* begins to form a customary part of the characters associated with tillage.

In a similar vein the book takes up the other specific verbs of agriculture: how to sow, reap, thresh, winnow and store. For each there is a developmental story, east and west. Seed drills were used in ancient Asia but appeared in Europe only in the 16th century. The large estates in Roman Gaul used machines to reap their wheat. But mechanical harvesting never entered China; even today mechanization is found chiefly in Manchuria, where they carry out wheat monoculture along American lines in large open fields. An enclosed-fan winnowing machine was used in Han times in China; here you can see a pottery model of the device, lately recovered from a grave. The most unexpected early invention is that of pelleting seed grain, using a pesticide held in gelatin. The recipe from the time of Christ is here: a strong soup of boiled horse bones laced with aconite flowers. Modern tests do not confirm that the method was worth the trouble it made for the sowers.

Crops are treated in some detail, not the grains alone but the full range, sources of the staple food *fan* and all its helper sauces *tsai*. The plants are figured from the wood blocks of the printed compendiums of Ming and Ching. The peanut is another post-Columbian introduction that has found large-scale success in China, mainly as an oil crop.

The volume ends with a bold and persuasive chapter that seeks to draw historical explanation from agronomical fact. Wet-rice production is open to steady gains in efficiency and productivity. Those improvements do not follow from input of heavy capital or from economy of scale. They arise from new seed, fertilizers, transplanting and weeding tools and the like. Skill is more important than equipment. The entire family works in the fields at the two labor peaks; in between sowing and reaping there is much less to do. Seasonal farm labor can hardly pay off, but the opportunity for home production, cottage industry of all kinds, is high. Family units prosper; industrialization is less favored. Laborsaving devices appear rather as modest gadgets than as factory machines, neither concentrating nor subdividing labor, nor demanding much capital. The South Chinese experts recommend still today that wet-rice fields remain smaller than a sixth of an acre, not much room for field machinery. Transplanting and weeding are skills no machines as yet hold.

The Chinese economy expanded for about eight centuries after the first "Green Revolution" under the Sung. The government of the time introduced a policy to enhance productivity so that they could raise large armies to defend the north and yet support a population that had largely crowded into the south away from the borders. Quick-ripening rice strains were brought in from Champa, a region now encompassed by South Vietnam, to allow double-cropping. New lands were watered, and trade was fostered in many countryside products. Long-distance transport of rice was improved, making possible the eventual growth of regional specialization, in sugar, silk, cotton, charcoal, lacquerware and many other products of skilled farm households.

Broad acres of dryland grain invite machine harvesters. Then the rural people are dispatched to work as mill hands in power-driven factories. Paddy rice instead feeds cities from a populous countryside world, its management unit the small-holding family. Wet-rice systems can support sustained economic development, but there landlords, merchants and even the state have access to the means of primary production only through the families at work on the land. "It is to the harsher dryland systems that we must look for real social transformation and historical *change*."

The argument is impressive, although not yet compelling. Little rototillers are noisily at work for rice-growing small holders in Japan; village collectives in South China grew plenty of rice, to be sure amidst turbulence from without; paddy rice is sown by aircraft across broad laser-leveled California wetlands. These are directions not yet fully considered; there must be other parameters in plenty. But it is all the same obvious that the symbiosis between human beings and their crops casts a subtle influence over the entire texture of human history. Agronomy is not so far from destiny.

MODERN BATTERIES: AN INTRODUCTION TO ELECTROCHEMICAL POWER SOURCES, by Colin A. Vincent, with Franco Bonino, Mario Lazzari and Bruno Scrosati. Edward Arnold, 300 North Charles Street, Baltimore, Md. 21201 (paperbound, $29.95). Nearly all the globe's automobiles start under the cranky power of the ubiquitous storage battery, a major product of the electrochemical industry. Those rechargeable lead-acid cells, properly described as lead–lead oxide cells with an acid electrolyte, go back to 1859, when Gaston Planté coiled two lead strips together around a linen spacer, put the assembly into a jar with dilute acid and charged up his new cell. A third of all the lead mined each year goes into 100 million such batteries, cheap, robust and now intensively engineered for easy maintenance. Their internal resistance is low through that conductive acid solution, and so power flows out freely, up to 100 watts per kilogram of cell, an output unmatched among practical batteries.

Do not expect that technology to produce your electric family car; true, lead-acid power density is only a few fold below that of a light automobile engine, but its energy density falls far short. A lead-acid cell able to store a kilowatt-hour weighs close to 100 pounds, whereas that same useful energy is stored in a pint of gasoline (along with the requisite gratis 10 pounds of ambient air). Forty thousand quiet little delivery vans for British milkmen, and many non-nuclear submarines with 180 tons of hefty battery in each, move slowly on lead-acid power. But only some new long-lived battery types can expect to compete with gasoline engines on the highway.

This well-written volume for nonspecialists is a "brief but comprehensive" engineering account of the gamut of batteries: how they work and how they are evolving. It modestly acclaims a current renaissance in battery development. During the past 15 years a bumper crop of batteries old and new has grown to nourish all those integrated circuits. One kilowatt-hour of electrical energy, costing at retail about 10 cents delivered by cord from a wall outlet, can be subdivided instead and fed to 5,000 perambulating wristwatches; it sells then for $10,000; quite a neat investment.

The leader of this Scottish-Milanese educational collaboration presents a long chapter of theory that makes pretty plain how a cell works. This is no easy task, and it has been shirked by textbooks. Take a strip of copper and one of zinc and place them in contact. It is easy to accept that the mobile electrons have a different binding energy within the two metals; why should they match? If that is true, then after contact electrons will freely diffuse between one metal and the other, until marginal readjustments offer no net energy difference.

At that point zinc has fed net electrons to the copper. They all reside in a surface layer a few atoms thick, in equilibrium between the atomic bonding and the mutual attraction of the electrons for the positive charge they left behind. The result is the formation of a double layer of charge, one layer with surplus electrons, one layer with an equal net positive charge. Between the two charge layers there is a strong field, but pass even a little distance beyond the double layer and the two equal and opposing thin layers in effect cancel. The new equilibrium is completed in nanoseconds after first contact, by the net migration of only one electron among each 10 million electrons freely able to move.

Every cell houses working double layers, equilibrium of which can be easily disturbed by draining excess electrons into an external conducting circuit. In cells double layers form, not at direct contact between two metals but at the interface between each conducting electrode and an ionic medium, the electrolyte. It is the flow of ions that brings in new charge to maintain the layer after disturbance. These enter from the other double layer some distance away, at a second interface between the electrolyte itself and a distinct electrode that provides the return for the external circuit. Ions leave the one electrode within the cell, to strike the best energy deal they can find. Other ions carrying the same net charge enter the complementary electrode; the external electron flow matches the inner charge exchange.

Net chemical reaction supplies primary energy as mobile ions and electrons move across the double layers and around the circuit. If all is right, the cell can be recharged by spending secondary outside energy to drive electrons inward, reversing the ionic flow and moving the charges back uphill to their initial sites at higher energy, thus restoring the entire chemistry da capo.

The potential difference is maximum only at the initial charge equilibrium; for any practical output a finite current must flow, diminished by frictional losses to heat and to various side reactions. The choice of cell reactions and their thermal and electrical characteristics, the wide variety of scale and purpose, ingenious engineering—particularly the more recent achievements—crowd chapter after chapter. There are intriguing devices, described category by exotic category, amid plenty of data and diagrams. The entire effort is supported by a good glossary and index.

The most used of primary cells, those called dry cells, are not dry at all. They have wet electrolyte, a water solution of ammonium and zinc chlorides, tightly sealed in, often gelled or absorbed in microporous separators. In about 1900 a couple of million of them were made and sold each year for flashlights and doorbells; today world production is up a thousandfold. In all these cells zinc metal oxidizes and manganese dioxide is reduced. (The usual carbon cathode is in fact only a passive collector of electrons.) Although two or three approximate accounts can be given, we still do not know in full detail what the discharge mechanism is for this cell, the reaction of which was proposed by Georges Leclanché in 1866. The preparation of the powdered manganese dioxide, a complex crystalline compound, is still ruled by subtle surface and impurity effects. Design of the Leclanché cell remains more an art than a science.

It is an ingenious art. The firm of Ray-O-Vac mass-produces flat dry batteries familiar in Polaroid automatic cameras. A vigorous nine-volt battery is about as big as your hand and three millimeters thick. Each of six layers within is a conductive sheet painted on one side with zinc powder and on the other with carbon black and manganese dioxide. The gelled electrolyte is held and absorbed by a thin synthetic netting that separates the stacked electrodes of this voltaic club sandwich; the whole is hot-sealed around the perimeter.

Two innovations will have to stand for two dozen discussed. It is evidently tempting to contrive a cell that is all solid state. Liquids and gels run and leak and corrode and are prone to form gases. Now, ions can of course drift through solids just as they diffuse through liquids. But in the usual ionic lattices ion transport is dependent on lattice defects, the charged atom hopping at random from one chance vacancy to the next. Defects are few at ordinary temperatures, and ions, thousands of times heavier than electrons, hop slowly. High ionic current through cold solids is hard to come by.

The most successful fully solid-state battery so far is a low-power lithium-iodine cell. The anode is solid lithium metal; the cathode is an organic iodine complex. The solid electrolyte is a thin layer of lithium iodide that forms at once on direct contact between anode and cathode. On discharge lithium ions diffuse across the lithium iodide, whereas lithium-ion vacancies migrate the other way. The electrolyte grows at the iodine interface, for heavy iodine ions migrate very slowly indeed. The internal resistance grows along with the thickness of the lithium iodide layer as the cell is discharged; supporting low-current electronics, a cell the size of a checker lasts reliably for up to 10 years. The cell finds most use in surgically implanted devices; cardiac pacemakers seem the right place for long-lived sealed batteries that cannot leak or bubble.

Quite another design current is flowing but is not yet in flood at the high-energy end. Massive storage of electrical energy held ready to even out the daily peaks and troughs of public power demand certainly pays off. Monster batteries present an alternative to ponded water and to turbines kept idly spinning in reserve. A few sturdy three-story warehouse buildings might house batteries enough for a city-scale load-leveling system able to store 100,000 kilowatt-hours almost anywhere. Such a device has the capacity of the power plants of three dozen nonnuclear submarines. Some battery!

High power density means large areas and good ionic conductors. But high energy density needs an atomic fuel that is not weighed down by many nucleons for every easily mobilized outer atomic electron. Hydrogen is the ultimate in lightness; its story, however, belongs to the budding technology of fuel cells, an endeavor that lies outside the scope of this book. Lithium, the next choice, is the star of the story, for both small powers and large. Almost a score of lithium-based cells are in wide use or just short of it.

A cheaper light metal, sodium, fuels the most practical new system at large scale. High electrolyte conductivity is essential, yet every aqueous or organic fluid reacts strongly with sodium. Molten salts might work. One solid electrolyte shows high promise; the substance is called sodium beta-alumina. Its crystal form is essentially the aluminum oxide lattice lodged between regularly spaced atomic planes of mobile sodium ions. For high ion conductivity its temperature must be kept above 300 degrees Celsius. The high-temperature power cell now favored has two molten electrodes, a central positive one of sulfur ringed by a negative one of sodium, the two melts separated by a ceramic cylinder made of sintered beta-alumina. During discharge sodium ions migrate through the hot solid wall to form sodium polysulfides within.

Such a beta-battery, as its developers call it, incorporates many steel-cased tubes, connected into modules. Once heated for start-up, a large battery will keep itself hot enough by means of its own thermal losses. Groups worldwide have developed this system into a reversible battery of good cycling life. It now yields three or four times the energy density of lead-acid batteries. For road use one must be prepared to control the consequences of accidents involving a red-hot battery box full of liquid sodium. For the calmer stationary duty of load leveling, tests of U.S. beta-batteries are now under way on the 100-ton scale.

ALBUM OF SCIENCE: ANTIQUITY AND THE MIDDLE AGES, by John E. Murdoch. Charles Scribner's Sons ($50). This excellent series seeks to convey the science of each period through its graphic representations. Professor Murdoch has marshaled his images mainly from the rich European holdings of the medieval manuscripts of the Latin West, thin booklike codices presented for the most part on parchment. He has by no means neglected the important Arabic and Persian material or the rarer papyrus scrolls of Egypt and the Levant. It is the Latinity, however, that is best preserved.

The science here is in large part the self-conscious preservation of and scholarly commentary on a few great classical authors, in particular the Greek mathematicians as well as Aristotle, Ptolemy and Galen. That pattern holds for Arabic and Latin scholarship alike, and sums up the lion's share of what we have. Discoveries, instrumentation, collections of material hardly figure. Science was *livresque:* it was "not just set down in books; it was largely carried out in books." The text was sovereign; even where it rested on empirical observation, once the theory inferred was written down, fresh appeal to the material itself rarely animated the continuators and commentators.

Copyists were trained for the task of writing, the tedious repetition of manifold permutations of few letters. It was hard enough, and costly enough, to perform that task well. The secondary task of getting pictures right lay outside that system. Talented specialists would have to be called on; a copyist could hardly be expected to be draftsman or painter. The bandwidth of illustration is too high; the very values it offers put it almost out of reach. It is no surprise to see model books, codices for the use of the trade, in which anatomical figures, foliage, diagrams from the theorems of Euclid and animals of various kinds are all provided for copying by illustrators. It is natural enough, but more of a surprise, to see a model drawing of the 13th century, here one of a stork, outlined in pinholes probably to allow quick copying by dusting with some powder. (No such use of the figure has been found.) Another page bears a specially drawn field, four bordered rectangles in a connected pattern, the frames left blank. The figure provided a logical format, useful for graphical explication of any number of textual arguments.

There are college textbooks, handbooks of simplification and synopsis, compact encyclopedic works meant for training and reference by clerics who would rarely need the full professional matter of the ancient masters. There are plenty of tables, mnemonic devices and systematic logical diagrams. The hand of the Venerable Bede was often drawn and elaborately annotated; his much-used eighth-century work on the reckoning of time gave explicit conventions for figuring with the fingers to derive the dates of a wide variety of religious feasts; it also offers a scheme for arithmetical finger reckoning up to 10,000.

Field guides serve physician and herbalist, surveyor and star clerk. Here one might expect more need to consult the objects of the world. A series of paintings of the wonderful herb mandrake is shown from medieval manuscripts. The version that went with an Arabic translation of the classical Greek text presents an unlikely stylized plant but not a magical one. Two more show the plant with roots that assume a magically human form: one still rootlike in texture and detail, the other about as good a drawing of the human body as its artist could make, topped by a wide foliage crown. The last version depicts a very believable but striking plant, its forked root real enough, although it still evokes the human form. Even this example was not drawn from a specimen plant; its prototype was realistically painted for a Paduan prince. One 15th-century artist illustrated his plants from the source by printing nature impressions made from real paint-coated leaves. Alas for his readers, he often misidentified the species.

All of this has a remarkably familiar look today, one that might have been missed even 10 years ago. These manuscript images are drawn from the same dwindled universe as today's computer graphics. Blank fields stand ready to accommodate some large class of data into a single neat format. Logical diagrams abound, often elaborated into trees and intricate flows. Realia, like flowers or animals or the human body, are stylized into conventional but recognizable outlines and shapes, easy to repeat (see Pac-man or any little red Apple). Simple transformations are much employed: the old manuscripts figured constellations twice, once as seen by looking out through the sphere of the sky and once by looking in.

The underlying reason is the same: copies are required fast. High resolution, any detailed attention to direct imaging from the variable and refractory outside world, takes too long. Today magnetic patterns are scarce goods, as paper was then, and copying too is simple, mainly symbolic. Presenting simplified images bit by bit is as much as either the old scribe or the modern microprocessor can generally do. Richly printed surfaces still offer more, perhaps through some powerful new form such as the video disk.

It is fascinating to see some fresh images. There is the painting of the Arabic library at Basra in the 12th century, the books neatly shelved; there are the varied exemplary bandages on arms and legs, teaching a medical art that appears to trace back long before the manuscript example, perhaps to the Hellenic original. Altogether the scholarly Harvard author-compiler has built us an exciting new vantage for looking at science's medieval past. Perhaps it is even a cautionary one. The sense that our quick digital descriptions are information-sparse compared with the dense thicket of real photons and atoms, that all theory is gray against the green world, is concededly a Romantic one. But is it outdated?

# IT ISN'T ESSENTIAL THAT ONE EXPERIENCE OTHER CARS BEFORE BUYING A SAAB. BUT IT CAN BE VERY CONVINCING.

The road to a Saab is littered with the hulks of cars you thought would be just what you wanted, but weren't.

When you wanted style, that's just what you got. You got tail fins, two-tones and flash. You got mundane engineering clothed in futuristic dazzle. A truck in a tuxedo.

When you wanted performance, you went out and bought it. You bought twice as many cylinders as necessary. Twice as many carburetor barrels. Twice as many exhaust pipes. You had to: you had to move around twice as much weight.

When you wanted economy or utility, you knew where to get it. You got it in a plain brown wrapper. No frills. No unpleasant surprises. No pleasant ones, either.

Whenever you wanted any of these things, you never had trouble getting them in a car. The trouble was, you could never get them all in the same car.

*The ideal car should do everything well.*

It probably isn't possible to make the ideal car. For one thing, especially in a big car company, it's often impossible to get everybody to agree on what the ideal car is.

Fortunately, Saab is a small car company. And all parties involved in the creation of Saab automobiles have come to the general agreement that, for their part, the ideal car is one that does everything well.

You can tell how close a carmaker has come to making the ideal car by examining that car's strong points.

Paradoxically, there really shouldn't be any.

In a Saab, for example, you'd be pressed to point out a single feature that is demonstrably superior to another.

| *1985 SAAB PRICE LIST** | |
|---|---|
| *900 3-door* | *$11,850* |
| *900 4-door* | *$12,170* |
| *900S 3-door* | *$15,040* |
| *900S 4-door* | *$15,510* |
| *Turbo 3-door* | *$18,150* |
| *Turbo 4-door* | *$18,620* |
| *Automatic transmission $400 additional.* | |

Saabs are generally acknowledged to be fine-handling cars. This is the result of front-wheel drive, rack-and-pinion steering, low-profile radial tires and excellent shock absorbers.

Yet a Saab's handling complements, rather than overshadows, its performance, which is equally impressive. That performance, in the case of the Saab 900 Turbo, is derived from the use of turbocharging, intercooling, 16 valves and dual-overhead camshafts.

Since there is no compelling reason why a Saab's performance and handling should not be perfectly compatible with comfort, room and utility, a Saab has those things too.

Is the ideal car, then, one that handles well, performs well and does many other things well besides? Well, ask yourself– why shouldn't a car do all this?

*A Saab could never be your first car.*

No one at Saab would be rash enough to claim they've made the ideal car.

On the other hand, no one at Saab would disagree that making the ideal car is their common goal. To understand this, is to begin to understand what a Saab is.

And many car buyers have begun to understand. Because, for the past four years, demand for Saabs has exceeded an ever-increasing supply.

Every year, for the past four, an increasing number of drivers have discovered that a Saab is close to their personal ideal of what the ideal car should be: fun to drive, practical, comfortable and durable.

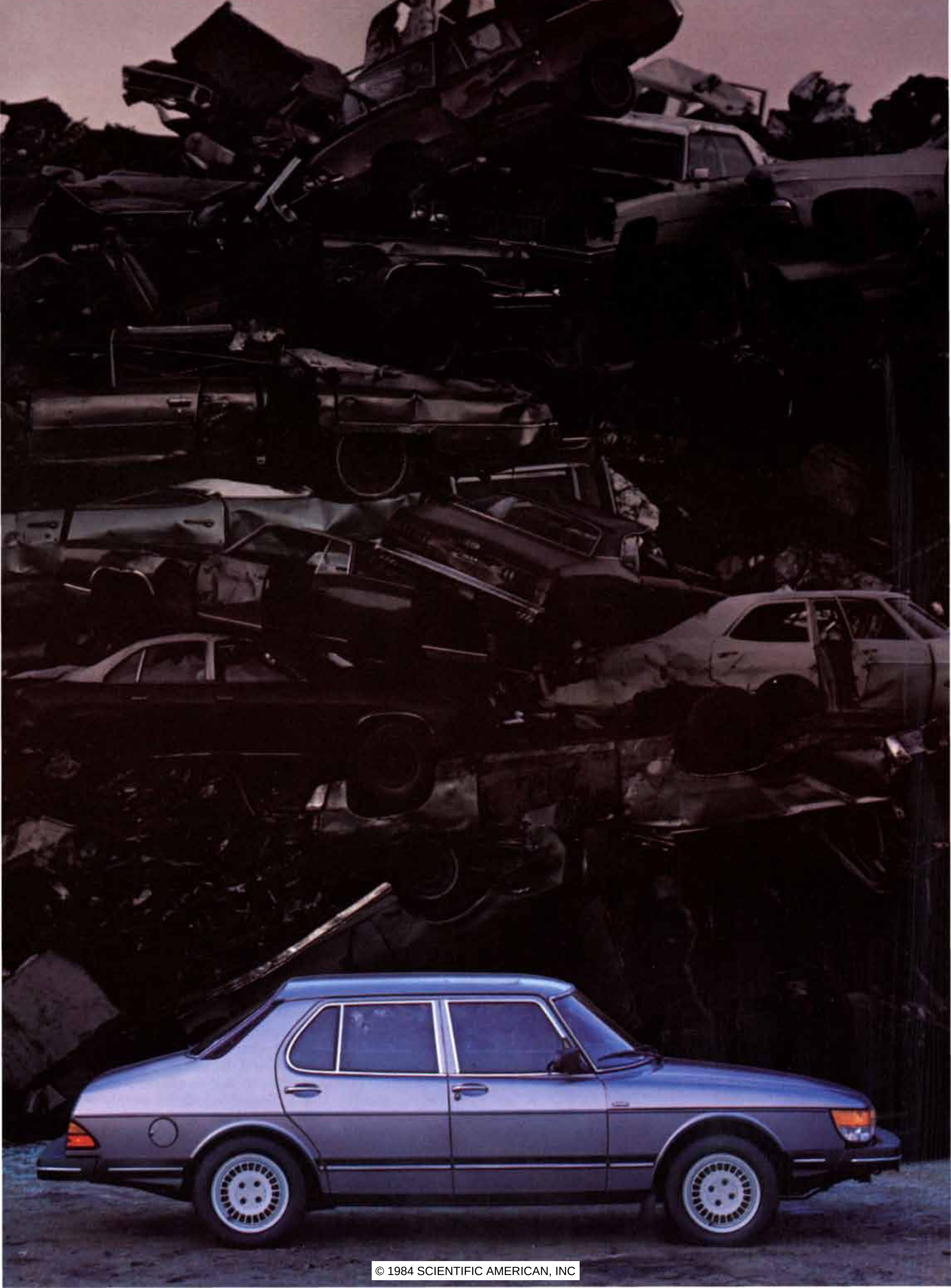
But it's doubtful they would have ever recognized all these Saab virtues without having lived through other cars' shortcomings.

Which is why it's a relatively safe assumption that a Saab could never be your first car.

Although it is highly likely it could be your last.

SAAB
*The most intelligent car ever built.*

**Manufacturer's suggested retail prices. Not including taxes, license, freight, dealer charges or options. There are a limited number of turbo models available with Saab's Exclusive Appointments group, which includes: leather upholstery, fog lights, and electric sunroof, at additional cost.*



Established 1845 SCIENTIFIC AMERICAN January 1985 Volume 252 Number 1

# The Command and Control of Nuclear War

*Nuclear weapons and strategic policy attract increasing public concern, but systems for command, control, communications and intelligence may be just as important in deterring nuclear attack and preventing escalation*

by Ashton B. Carter

Weaponry tends to dominate discussions of nuclear war. Missiles and bombers, throw weights and flight times, and the elaborate counting rules of arms-control agreements provide the grist of public debate. After the weapons themselves come the plans for their use, bearing such names as minimum deterrence, flexible response and countervailing strategy. Weapons and strategic doctrine are meaningless, however, unless the superpowers also have the means to know what is happening in the chaos of crisis or war, to provide for decisions by legitimate authorities and to have orders carried out precisely and faithfully. In military parlance these capabilities form the system of strategic command, control, communications and intelligence, or $C^3I$ (pronounced "see cubed eye"). Although $C^3I$ has been largely neglected outside a narrow circle of experts, it is an all-important facet of the problem of deterring nuclear war, fully as important as weapons and doctrine.

All concepts of nuclear war have built into them important assumptions about $C^3I$. For example, deterrence presupposes that the nation attacked can communicate retaliatory orders to its weapons in spite of the destruction of its national capital and normal communications facilities. The idea that retaliation should be appropriate to the attack presupposes that national leaders would in fact have a clear idea of the extent of the damage inflicted; it further assumes that the counterattack would resemble what the leaders had in mind and that its "appropriateness" would be recognized by the enemy leadership.

Scenarios describing how a "limited" nuclear war might be fought imply that escalation from less than total to total nuclear war is not automatic and that limited wars can therefore be waged purposefully and coherently. The concept of protracted nuclear war levies an even stronger requirement than a single exchange of warheads does: coherent command, control and communication must persist for weeks or months after a major nuclear attack. According to Secretary of Defense Caspar W. Weinberger's Annual Statement for Fiscal Year 1983, the U.S. must possess the means "to impose termination of a major war on terms favorable to the United States and our allies even if nuclear weapons have been used." Yet U.S. military leaders have also suggested the U.S. might attack the Soviet leadership. If this vital element of the $C^3I$ system of the U.S.S.R. were to be eliminated, who would be available to cooperate in terminating the war?

Such considerations raise the question of whether the $C^3I$ problem renders futile many prevailing theories about nuclear war and plans for using new weapons. Some nuclear strategists imagine that a nuclear war would unfold like a chess game. Chess players, however, have complete knowledge of the positions of all the pieces, can execute precisely every move they want to make and can work out the possible consequences of each move. An analogous clarity in nuclear war is most unlikely. In addition to its importance for the deterrence of nuclear war, $C^3I$ forces planners to think through the potential course of such a war in vivid detail. Consideration of $C^3I$ thus lends a needed concreteness to the abstractions of nuclear strategy.

Until a few years ago public discussion of $C^3I$ by knowledgeable Government officials was rare, since the issue was rightly regarded as extremely sensitive. This situation began to change, apparently because many officials believed that without more prominence $C^3I$ would never get the attention it deserved. Consequently in recent years analysts have explicated the $C^3I$ problem in lurid detail, even putting forward a number of alarming possibilities: the U.S. is a paper tiger that could not in fact retaliate after a nuclear attack (since the command structure could be "decapitated"); reliance on a strategy of launch under attack for ICBM's could invite disaster if warning sensors mistakenly indicated a Soviet attack; electromagnetic-pulse (EMP) effects could disrupt so much electronic equipment that most communications and computer systems simply would not work. Less common at this stage of the evolution of the $C^3I$ issue is constructive advice about what can be done to make these probabilities smaller. And looming behind the quest for solutions to $C^3I$ problems is the unsettling suggestion that no matter how good the C³I equipment is and how well trained its users are, an ineradicable residue of uncertainty will always remain about the unprecedented circumstance of nuclear war.

As C³I has begun to enter public discussion it has also begun to receive renewed emphasis in actual planning. The Reagan Administration has made C³I a key aspect of its strategic modernization program, which includes the MX intercontinental ballistic missile (ICBM), the Trident II submarine-launched ballistic missile (SLBM), sea-launched cruise missiles (SLCM's), the B-1 and Stealth bombers, air-launched cruise missiles (ALCM's) and the Strategic Defense Initiative, or "Star Wars" antimissile defense research plan. C³I, however, still accounts for a small part of the budget spent on strategic forces. About a dime out of every dollar allocated to strategic forces goes to C³I, which means somewhat more than a penny out of every dollar in the defense budget.

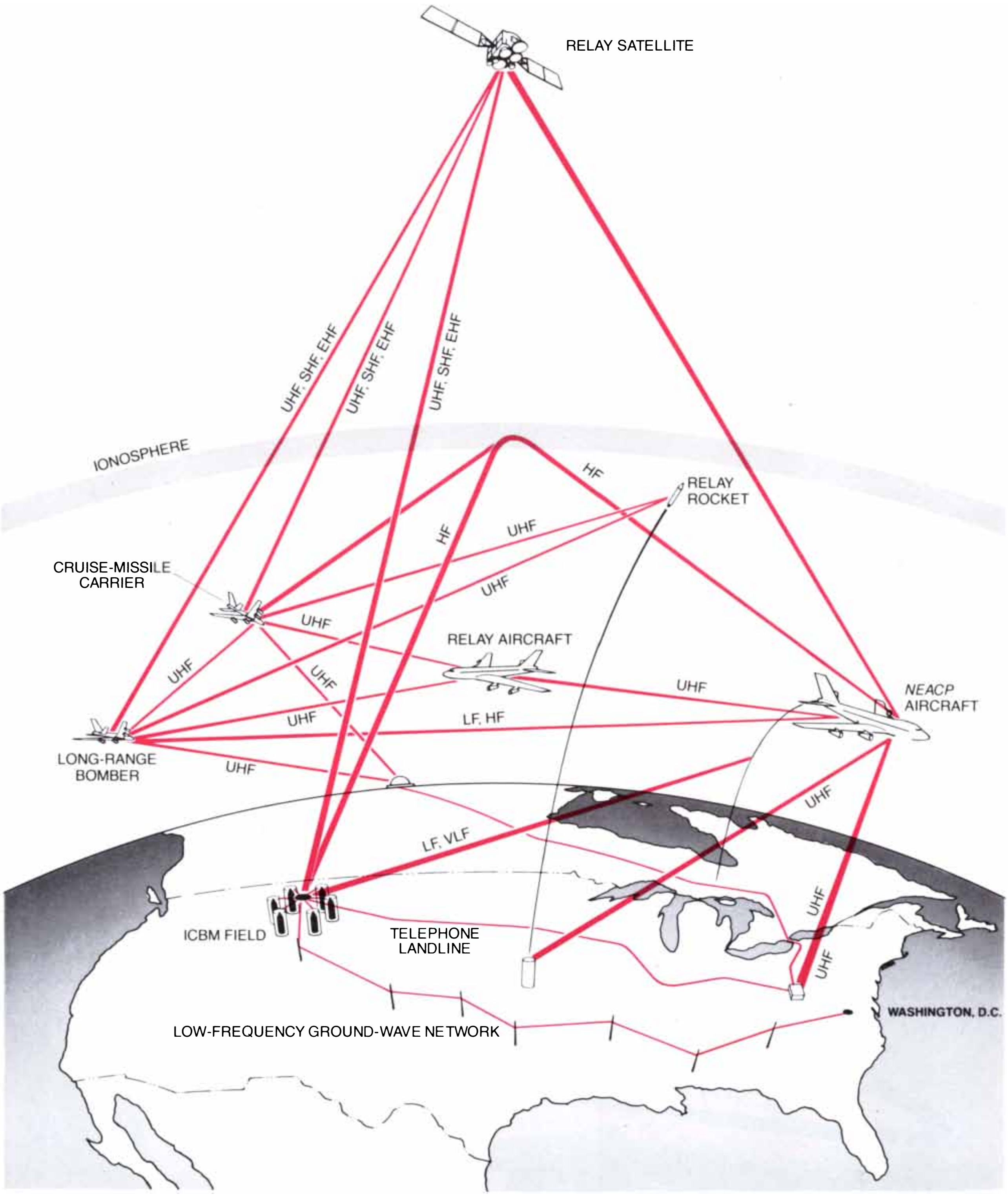
Concretely, the C³I system consists of four parts: command posts, sensors, communications links and procedures for the use of all this equipment. Command posts are needed to keep national leaders alive and in touch with the situation. Sensors provide both "strategic" warning, indicating imminent Soviet attack, and "tactical" warning, indicating that attack is already under way. The collection of strategic-warning data is allied with peacetime intelligence collection; this aspect of the "I" in C³I will not be discussed further here. In addition to sounding the alarm, warning sensors can also record for those who make decisions (and for history) where the bombs fell. Communications links carry warning data from sensors to command posts and orders from command posts to the nuclear forces. Procedures need to be worked out well in advance to rescue the leadership from Washington. Procedures must also be devised to reconcile the military's duty to exercise negative control over nuclear weapons ("Do not shoot until told") with positive control ("But respond reliably to authentic orders"). Nuclear forces must be managed in crises without unwanted provocations. Even such minutiae as tuning to a common radio frequency must be prescribed in advance. This article deals mostly with the first three elements of the C³I system, since they present well-defined technical problems and opportunities for improvement.

In approaching this task it should be kept in mind that even elaborately engineered systems can produce catastrophic results when they are faced with unexpected circumstances and pressures. In a celebrated incident in the summer of 1980 a faulty component in a data processor at the command post of the North American Aerospace Defense Command (NORAD) in Cheyenne Mountain, Colo., began generating spurious warnings of a Soviet missile attack. This incident was unimportant in itself, since the spurious data stream did not simulate a plausible attack. (The signals changed erratically with time, and the international situation was calm.) All participants in the alert agreed that the U.S. came nowhere near to "accidental war." A less easily recognized malfunction in tenser circumstances, however, could be more troublesome.

Furthermore, military operations are rife with opportunities for C³I foul-ups. The likelihood that something will go awry is compounded by the fact that the military system is so vast and complex that senior officials cannot oversee its every move. Moreover, since no one has ever been in a nuclear war, leaders would have to rely largely on preconceived notions rather than on experience and reflection. A healthy regard for Murphy's law is probably the better part of wisdom in strategic C³I.

The first technical ingredient of any strategic C³I system is a set of survivable command posts. A command post is not just a "Führer bunker" in which to hide the national leaders. They would be helpless without warning-sensor displays, communications

**PRESIDENTIAL "DOOMSDAY" PLANE, officially known as the National Emergency Airborne Command Post, or NEACP (pronounced "kneecap"), is one of a fleet of specially modified Boeing 747's maintained in a high state of readiness by the U.S. Air Force for possible use in a nuclear war. In an emergency the National Command Authorities—the president and the secretary of defense (or their duly deputized alternates or successors)—would board the plane, which would then fly to its cruise altitude, where it would be in a position to communicate with the nation's strategic forces by various line-of-sight techniques (*see illustration on next page*). The plane is also capable of transmitting low-frequency (LF) and very-low-frequency (VLF) signals over the horizon by means of a long antenna trailed from the back of the plane. In addition each NEACP plane carries a satellite-communications terminal and an assortment of other gear. Although there is no way at present for the Soviet nuclear forces to target such an airborne command post once it is in the air, the plane is considered quite vulnerable to a surprise submarine-launched missile attack when it is on alert on its airstrip.**




**VARIETY OF COMMUNICATIONS LINKS** are available to enable U.S. leaders to stay in touch with the nation's nuclear forces in an emergency. In this idealized scene the NEACP aircraft has already taken off from an air base near Washington, D.C., following the receipt of an early warning of an impending missile strike by the U.S.S.R. The aircraft is shown in communication with two components of the U.S. nuclear forces: a field of Minuteman strategic missiles in Montana and a detachment of long-range bombers and cruise-missile carriers that have taken off from their bases and flown north over Canada, where they await the order to proceed to their targets in the U.S.S.R. (If they receive no message, they will automatically return to their bases.) Because each method of communication has its own drawbacks, the system must be designed with several redundant links. For example, since telephone landlines could be severed in a nuclear attack, satellites and relay aircraft would be used to provide line-of-sight communications links over the horizon. Satellite signals, on the other hand, could be disrupted by high-altitude nuclear explosions; the resulting "blackout" effect would be particularly severe at lower frequencies. Furthermore, nuclear explosions could inject enough extra electrons into the lower ionosphere to cause high-frequency (HF) signals to be absorbed rather than refracted back down to the earth. Broadcasting at lower frequencies would require that the NEACP plane reel out a transmitting antenna several miles long. An alternative communications method calls for reading the presidential command into a tape recorder and launching the recorder on a rocket high enough to have line-of-sight contact with the strategic forces. Not all the technical possibilities shown have actually been deployed.

terminals, codes for preparing appropriate orders and a highly trained staff conversant in the arcana of fighting a nuclear war. Strategic command posts therefore must be substantial facilities, providing much more than just physical protection.

The U.S. has considered the entire spectrum of technical possibilities for survivable command posts, from deep-underground war rooms and mobile surface vehicles to ships and submarines, but it has acknowledged particular reliance on airplanes. The Air Force maintains a fleet of specially modified Boeing 747's known as National Emergency Airborne Command Posts, or NEACP's (pronounced "kneecaps"). Once such a plane is in the air it cannot be targeted. By flying at high altitude it can communicate effectively by various line-of-sight techniques over a wide area, and it can transmit very-low-frequency (VLF) signals over the horizon by means of a long trailing-wire antenna. Each NEACP plane also carries satellite-communications terminals, warning-data receivers, EMP detectors and an assortment of other gear, as well as seats for the presidential entourage.

Although a command post of this type is comparatively safe when it is airborne, it is quite vulnerable when on alert on its airstrip. An SLBM fired from a submarine just off the coast of the U.S. could arrive over the air base within 10 minutes of its launching; since it would take several minutes for the crew to get into the plane, bring the engines up to speed, taxi to the runway and take off, the ability of NEACP to survive a surprise attack depends on extreme vigilance.

One safeguard against such a surprise attack is another fleet of airborne command posts, code-named Looking Glass and operated by the Strategic Air Command. Since 1961 at least one Looking Glass plane with an Air Force general on board has been in the air at all times, 24 hours a day, 365 days a year. Presumably some of the NEACP fleet would also take to the air in a time of crisis.

Notwithstanding the fact that NEACP and Looking Glass cannot be targeted in flight, there are problems inherent in relying excessively on aircraft for survivable command posts. Even in flight such command posts are vulnerable to some nuclear-weapon effects, such as EMP, radioactive clouds, turbulent air from thousands of fireballs rising all over the country and harmful dust inhaled by their jet engines. Without aerial refueling or a place to land and refuel, the president could not wait long to make decisions involving communications by means that require NEACP to be airborne, such as line-of-sight techniques or VLF trailing-wire antenna. New airframe designs may help to alleviate this problem by making it possible to deploy aircraft that can "loiter" aloft for longer periods. Looking to the future, emerging technologies could enable the U.S.S.R. to track NEACP and give in-flight commands to missiles to home in on its position.

Clearly a vital question about a national command post is "Who is to be in it?" U.S. officials understandably avoid public comment about the sensitive issue of how continuity of government is to be ensured in a nuclear war. The authority to order the use of nuclear weapons is lodged formally with an entity called the National Command Authorities, which under normal circumstances consists of the president and the secretary of defense. Lines of succession for these offices are established, and so in theory the body would continue to exist even after the death of the principal officeholders. How the U.S. would in fact ensure the survival of an authorized command entity in the event of a nuclear war is a matter of top-secret procedures worked out in peacetime.

Warning sensors are the second major category of C³I equipment. There is very little the U.S. could do in the 10-minute flight time of an SLBM or the 30-minute flight time of a land-based ICBM to prepare American society for a nuclear attack. Nevertheless, missile-warning sensors do serve certain crucial strategic functions. Bombers, cruise-missile-carrying aircraft and airborne command posts all need immediate warning to enable them to escape from their bases before SLBM warheads begin to explode above them. If a launch-under-attack threat is to be credible, the U.S. must show that it can reliably receive early and accurate evidence of an attack by the U.S.S.R. The most important use of the data from early-warning and attack-assessment sensors could well be the record such information provides the president about the nature and scale of the attack on the U.S. Without this information, which is not likely to be readily obtained after the warheads have fallen, the president cannot choose an appropriate response.

As in the case of command posts, there is a wide variety of technical possibilities for early-warning and attack-assessment sensors. It would be reassuring to deploy a number of sensors operating according to different physical principles, since their output is to be put to such a momentous purpose. The sensors must obviously survive long enough to do the job, as must the communications links from sensors to command posts.

A novel attack-assessment system called the Integrated Operational Nuclear Detection System (IONDS) is scheduled to be launched by the U.S. in the late 1980's aboard the NAVSTAR navigational satellites of the Global Positioning System (GPS). The IONDS sensor package will include visible-light sensors, X-ray sensors and EMP sensors to detect nuclear explosions in the atmosphere and in space. A nuclear explosion in the atmosphere produces a characteristic double-peaked light pulse with a structure that depends on the explosive yield of the nuclear warhead. By measuring the time of flight of the flash to several satellites it should be possible to locate each burst to within a fraction of a kilometer. The IONDS data would serve not only to characterize an attack on the U.S. but also to verify that the U.S. warheads launched in retaliation had reached their targets in the U.S.S.R.

In addition IONDS would record the detonation of Soviet SLBM's on U.S. territory some 10 to 20 minutes before the expected arrival of the more accurate "silo killing" Soviet ICBM's. The U.S. leadership would have additional information to use in making the dangerous decision of whether to save the threatened ICBM's by launching them promptly. NAVSTAR satellites will incorporate features designed to increase their resistance to nuclear-weapons effects and also to possible future Soviet antisatellite weapons.

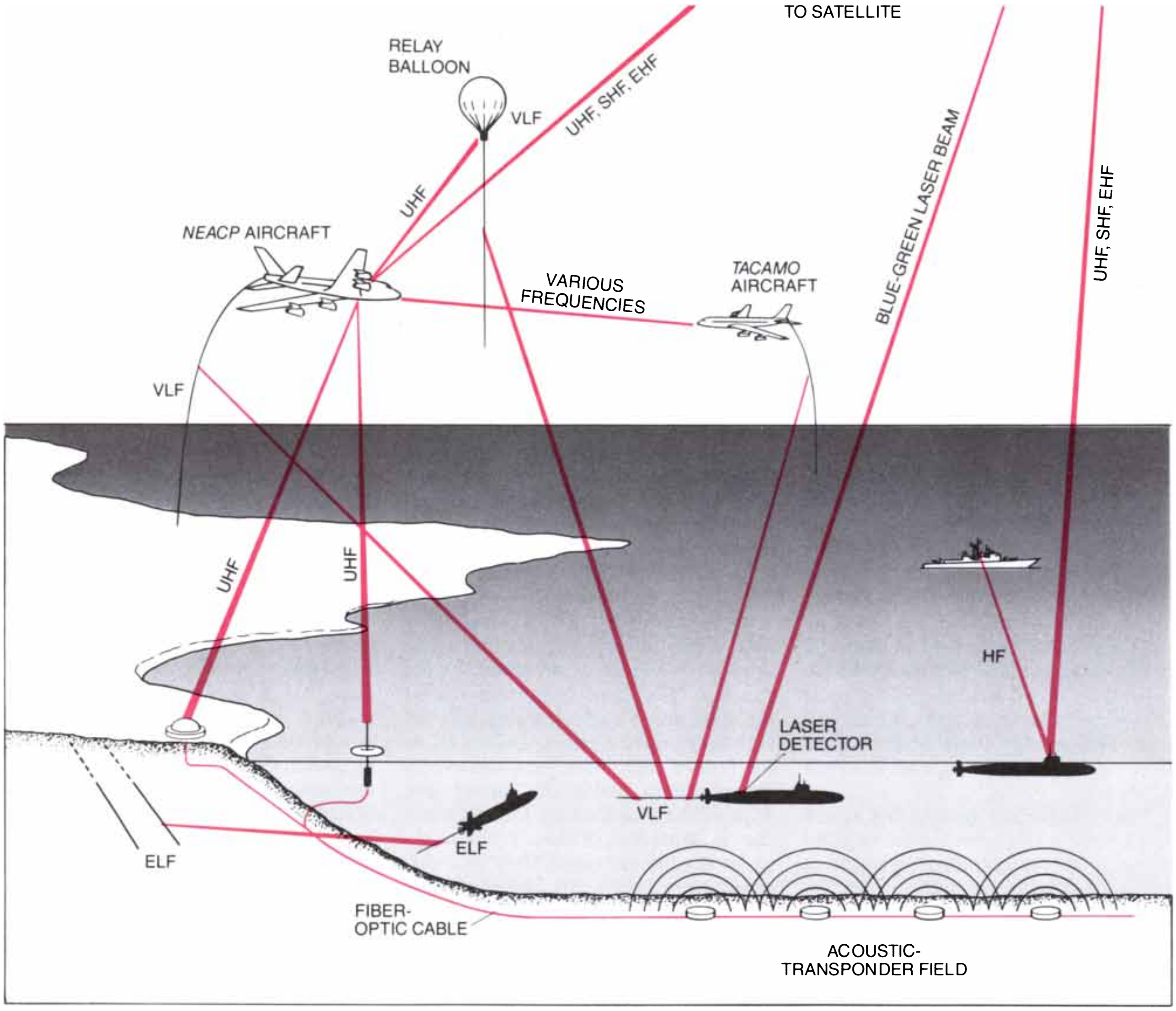
Each kind of nuclear force has distinctive communications requirements. For example, the C³I system must be capable at a minimum of transmitting launching orders to the Minuteman ICBM fields in the Middle West from Washington or from some other over-the-horizon location. The Emergency Action Message, or "go code," is a short, preformatted, encrypted message. More demanding would be the rapid response needed to launch the U.S. ICBM's that were under attack before the Soviet ICBM warheads arrived to destroy them in their hardened silos. Some strategic planners have suggested that any missiles surviving an attack should be retargeted to ensure that the highest-priority targets in the U.S.S.R. would be covered in the retaliatory strike. Such retargeting would impose yet another sophisticated communications demand.

Bombers and cruise-missile-carrying aircraft have even more complex needs than ICBM's. Rapid and reliable communications from missile-warning sensors are necessary to enable the aircraft to take off and avoid destruction on the runway. Once they are airborne, the planes would head north to predesignated locations, where they would loiter aloft and await an Emergency Action Message. If they received no order to continue on to their targets, they would return to their bases. It is not clear whether bomber crews that had carried out their bombing runs over the U.S.S.R. could expect to land on the territory of U.S. allies near the U.S.S.R., refuel and return to whatever airfields remained in the U.S.; if they could, they would need additional $C^3I$ facilities to do so.

The task of communicating to missile-carrying submarines is complicated by the long distances involved, by the paramount need for the submarine to remain hidden and by the opacity of seawater to all electromagnetic radiation except extremely-low-frequency (ELF) and VLF radio waves and blue-green light. On the other hand, the survivable submarine force would not need to depend on rapid communications or reliable warning to do its job.

In peacetime U.S. missile submarines are normally tuned to VLF broadcasts from large land-based transmitters, which could be readily destroyed at the outbreak of a nuclear war. In wartime the U.S. would rely heavily on an airborne relay system named TACAMO (for "Take charge and move out"). The TACAMO plane receives an Emergency Action Message from the national leadership and rebroadcasts it over a VLF antenna several miles long, which it trails from its tail. The submerged submarine must trail a long receiving antenna a few meters or less below the ocean surface, within the seawater penetration depth of VLF radio waves. Although the submarine itself can remain deeper than the antenna, it still must limit its depth and speed if the antenna is to work properly.

Proposals exist to exploit the two other seawater "windows," which lie in the ELF and blue-green parts of the spectrum. A radio antenna operating in the ELF range (between 70 and 80 hertz) is known to excite the character-




**COMMUNICATIONS WITH MISSILE SUBMARINES are limited by the fact that seawater is transparent to electromagnetic radiation in only three parts of the spectrum: at extremely low frequencies (ELF) and very low frequencies (VLF) in the radio region and at blue-green frequencies in the visible region. Radio broadcasting at the long wavelengths characteristic of ELF and VLF transmission has two disadvantages: it requires very large antennas and power supplies, and it is limited to very low data rates. On the other hand, radio waves at these frequencies would propagate well over long distances even if the ionosphere were to be disturbed by nuclear explosions. The U.S. Navy now relies primarily on VLF relay aircraft code-named TACAMO (for "Take charge and move out").**

istic modes of vibration of the resonant cavity formed by the earth's surface and the ionosphere. This phenomenon makes it possible for an ELF signal to propagate worldwide. Since the depth to which an electromagnetic signal is able to penetrate a conducting medium such as seawater increases with the wavelength, a deeply submerged submarine can in principle receive such an ELF signal.

For the sake of efficiency, however, the diameter of a transmitting antenna must be approximately equal to the wavelength of the radiation it broadcasts. Its data rate in bits per second is usually limited to a fraction of its carrier frequency in hertz. Hence an ELF antenna for a system with an effective data rate of only a few bits per minute must nonetheless be many miles in diameter. One such system has been built by the U.S. Navy in Wisconsin; its main disadvantage is that it is extremely vulnerable to nuclear attack.

Another communications scheme involves blue-green laser light beamed from a satellite to sensitive detectors mounted on the hull of the submerged submarine or towed closer to the surface. This idea is still in the research stage, and its usefulness is uncertain. Such a communications system would obviously depend on the satellite's ability to survive a direct attack.

It is possible that long VLF antennas trailed from balloons could supplement or replace TACAMO aircraft in certain circumstances. If the submarine were to tow a small buoy with an antenna, it could receive an Emergency Action Message by satellite or by high-frequency (HF) radio relay from other Navy ships at sea. Still another possibility involves sowing areas of the ocean floor with acoustic beacons linked by fiber-optic cables to floating buoys or ground stations. Altogether there is quite a rich array of technical possibilities for postattack submarine communications and no fundamental reason why an appropriate ensemble of these links should not be as reliable as the postattack links to ICBM's.

A strategic communications system must be designed to resist all kinds of interference, including physical destruction, jamming, interception or mimicking by enemy intelligence, disturbance by the ionosphere and disruption by the EMP effect. It must be assumed that radio antennas, land-line switching centers and ground-based satellite-control facilities would all be targeted in a major nuclear attack. Satellites can malfunction or fail completely without periodic "housekeeping" signals sent from ground-based control stations. As a consequence the U.S. military is planning a new generation of autonomous spacecraft that will generate most of their own control instructions with the aid of an onboard computer; in addition control facilities will be deployed in trucks that would roam the nation's highways to avoid being targeted.

Antisatellite weapons present a special kind of threat to strategic communications systems. For example, the U.S.S.R. could in the future plan to attack U.S. communications satellites with interceptor missiles, laser weapons, particle-beam weapons or remote-controlled "space mines" placed in orbit next to U.S. satellites [see "Antisatellite Weapons," by Richard L. Garwin, Kurt Gottfried and Donald L. Hafner; SCIENTIFIC AMERICAN, June, 1984]. A nuclear explosion in space could produce enough X-radiation and gamma radiation to damage a satellite at a range of many hundreds of kilometers.

Jamming a radio link means transmitting "noise" to the receiving antenna in order to drown out meaningful signals from "friendly" transmitters. A distant transmitter can jam effectively at high frequencies and very low frequencies, since these signals can propagate over the horizon. For the highest frequencies, however, the jammer must have a clear line of sight to the receiver being jammed. Operating in this way, Soviet forces in Cuba or on board ships off the U.S. coast could seek to jam the "uplinks" to satellites positioned over the U.S.

The most direct way to thwart jamming would be to increase the power of the friendly transmitter. Another way would be to use directional antennas. A directional transmitter could focus the radiated energy toward the receiver; a directional receiver would readily accept energy from the friendly transmitter. Because the directional gain of a satellite antenna depends on the ratio of the signal wavelength to the antenna diameter, extremely-high-frequency (EHF) and superhigh-frequency (SHF) communications links would be easier to protect with directional antennas than the longer-wavelength ultrahigh-frequency (UHF) links in widespread military use today.

Radio transmissions typically occupy a narrow range of frequencies called the bandwidth. A jammer must radiate noise throughout much of this bandwidth or enough of the message may reach the receiver through the unjammed part of the band to make it intelligible. Accordingly another way for the sender to improve jamming protection is to increase the bandwidth of the signal. Since the available bandwidth increases with increasing frequency, EHF satellite links would again be superior to UHF.

Still another way to protect against jamming is to lower the data rate, in effect repeating the message several times to ensure reception. Since such methods use a larger bandwidth to support the data rate than would be needed in the absence of jamming, they are called spread-spectrum techniques. The repetition pattern must be kept secret or the hostile jammer could concentrate his jamming effort in just those time and frequency intervals that contain a crucial part of the message.

Spread-spectrum antijamming techniques are closely related to methods used to encode messages. Communications security is necessary to prevent the enemy from listening in to communications among U.S. leaders or sending false messages to U.S. forces. Spread-spectrum techniques can also be used to spread a low-power transmission over such a wide bandwidth that the enemy, ignorant of the pattern needed to process the message, cannot ferret the signal out of the background noise. Using this technique with satellite communications at extremely high frequencies, submarines might in the future be able to communicate back to shore without having their transmissions detected and located.

A nuclear war would present some special communications problems. Ordinarily HF radio signals pass through the lower ionosphere, where the electron density is low, and are refracted back to the earth by the electron-density gradient in the upper ionosphere. High-altitude nuclear bursts and the radioactive clouds from lower-altitude bursts would increase the electron density in the lower layer of the ionosphere. HF signals passing through the ionized lower layer would then be absorbed, since in this layer there would be a high density of ions and neutral atoms. This "blackout" of HF signals could last for long periods.

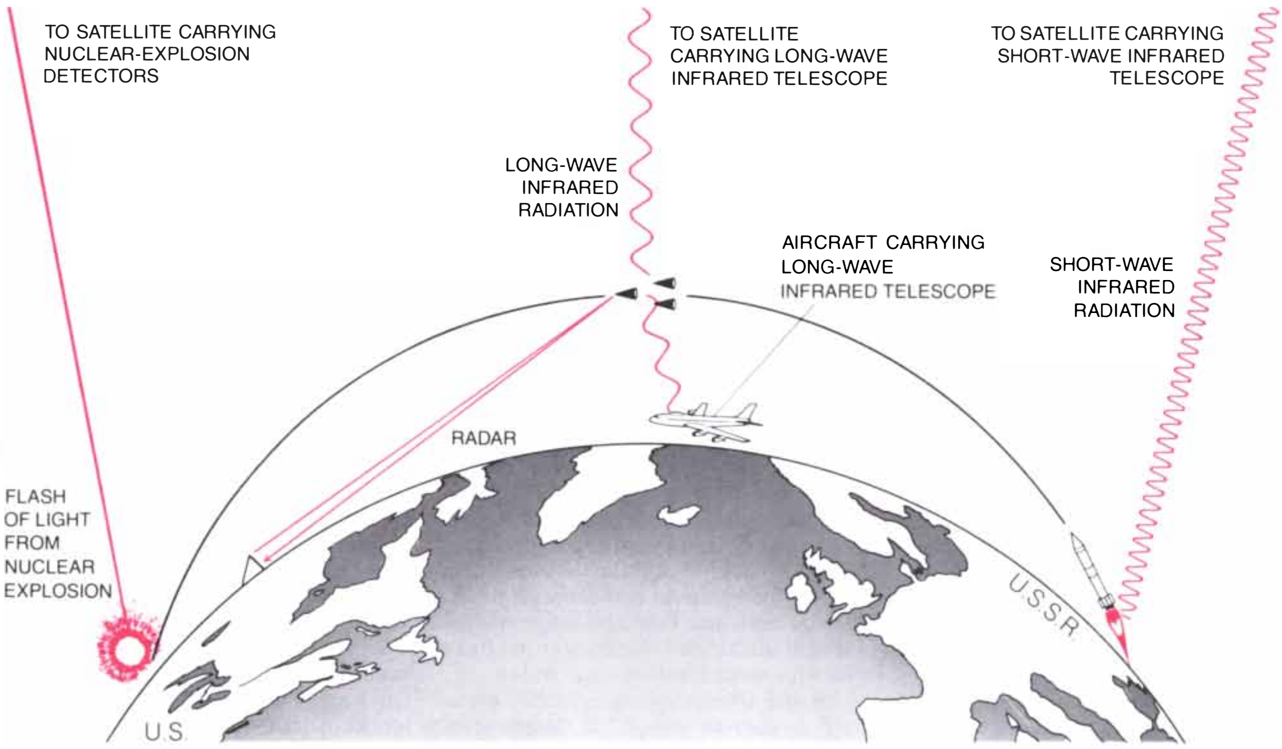
UHF satellite signals and VHF and UHF radar beams passing through severely ionized regions could also suffer absorption. Since the absorption coefficient is inversely proportional to the square of the radio frequency, EHF and SHF signals would suffer far less from blackout than UHF signals. Nevertheless, even these frequencies could be subject to the transient interruptions called scintillations, caused as parts of the signal wave front that pass through regions of different electron density (and therefore undergo different phase shifts) interfere at the receiver.

Electromagnetic pulse is a well-publicized form of interference. It is generated when gamma rays from a high-altitude nuclear explosion induce strong electric currents in the upper atmosphere. The resulting intense radiated fields would contain frequency components ranging from ELF all the way up to VHF. The EMP effect from a single burst at an altitude of several hundred kilometers could blanket the entire U.S. Such a pulse would enter electronic equipment through apertures and along power lines and other conductors, causing potentially harmful voltage surges. Although this effect has been analyzed intensively, complex electronic systems have so many possible failure modes, each differing in many particulars from others, that predictions are hard to make.

The many potential vulnerabilities of the C³I systems for the support of strategic nuclear forces demonstrate that it is difficult to guarantee that the U.S. could carry out the most rudimentary aspect of its nuclear deterrent policy: to discern the nature of an attack by the U.S.S.R. and to retaliate according to a prearranged plan. More detailed doctrines for conducting possible nuclear wars require correspondingly more ambitious C³I systems. For example, strategists who foresee a nuclear war beginning with a less than total exchange and continuing through further exchanges must also envision a C³I system capable of supporting repeated cycles of attack and counterattack. These strategists must assume, for example, that attack-assessment sensors damaged in a first strike could still collect information and transmit it to command posts to enable leaders to make what the strategists prescribe as the appropriate response to the second attack. In a single exchange, in contrast, the warning systems might still be undamaged at the time they had to do their job. One positive note is that the situations in which the U.S. would want to make a limited response to a Soviet attack would probably be those where the attack was itself limited, and hence where damage to the U.S. C³I system was less than total.

If the war were to be protracted, continuing "tit for tat" for several weeks or months, new problems would arise: bombers and airborne command posts would have to find surviving airfields at which to land, refuel and prepare for the next round; generators and batteries used to power support equipment in the ICBM silos after power lines were down would go dead, and satellites would be lost without ground-control commands. At the other end of the time spectrum of possible nuclear wars lies a launch of ICBM's under attack. Such a move would require near-perfect confidence in warning sensors, qualified leaders on hand at the moment of attack and rapid transmission of launching orders over the communications links.

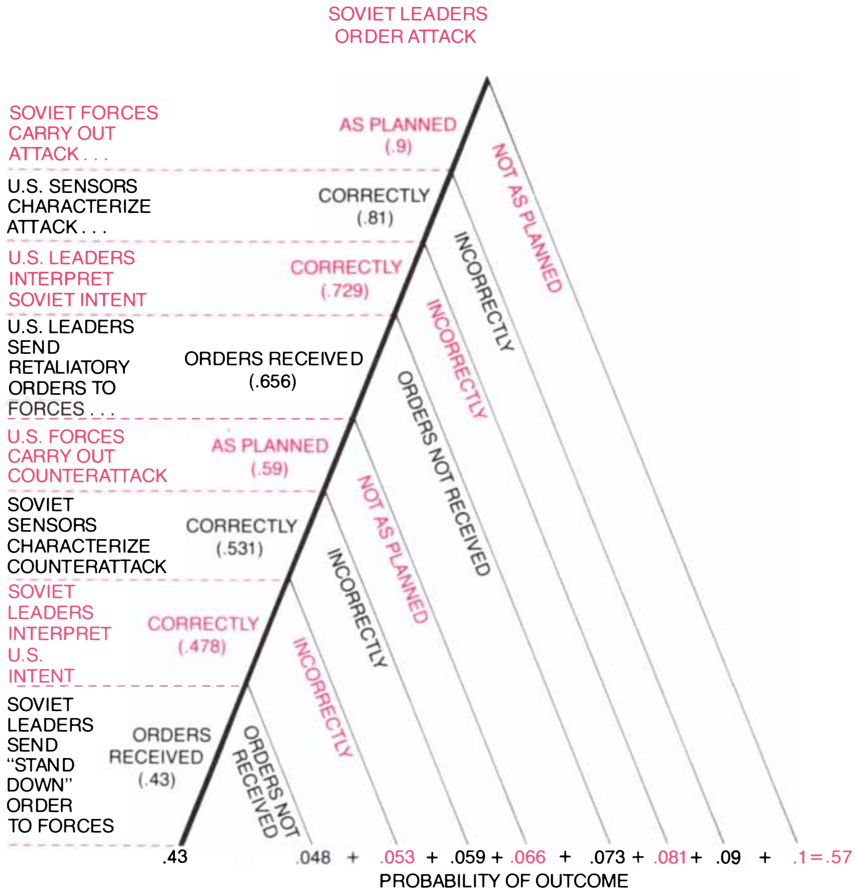
When one turns from purely military C³I to the broader needs of governments in responding to the desperate circumstances of nuclear war, one encounters further complications. The coordination of civil defense and recovery efforts, perhaps hopeless in any event, could only be made harder by the destruction of the nation's communications infrastructure. Deciding to use the nuclear weapons deployed in Europe might involve a complicated conference of the leaders of the North Atlantic Treaty Organization (NATO). Such a conference is not likely to be




**VARIETY OF SENSORS are available to provide early warning of an intercontinental-ballistic-missile (ICBM) attack and to assess the nature of the attack. U.S. leaders would presumably respond to such a warning by ordering airborne command posts, bombers and cruise-missile carriers into the air before they could be destroyed on the runway. They might also decide to give the launch-under-attack order to U.S. ICBM's before they could be destroyed in their silos. Perhaps the most important use of the warning data, however, would be made after the attacking missiles had arrived at their targets: since the output of the sensors might be the only clear indication the president ever gets about the scale and intent of a nuclear attack on the U.S., this information could be instrumental in determining whether and how the U.S. would retaliate. Short-wave infrared sensors would first detect the hot exhaust plumes of the missiles. Radars would later observe the approaching warheads and missile fragments. Later still, other satellite-based sensors would detect the actual nuclear explosions. Long-wave infrared sensors, which could record the infrared glow of the warm warheads against the cold background of space, have not been deployed, but they and other advanced sensors are currently being developed by the U.S.**

any easier to arrange than communications within the more cohesive U.S. military, which would be hard enough in the circumstances.

The leaders of the U.S. and the U.S.S.R. would communicate, after a fashion, through the violent actions they ordered. Beyond that, it is unclear how they could confer in order to bring hostilities to an end. In peacetime the hot line serves this purpose. This teletype system linking the capitals of the two countries is being upgraded to include facsimile transmission, enabling the leaders to trade charts and maps. Proposals to add voice or video links have been consistently turned down for fear that such intimate, real-time contact between leaders in a crisis could foster subjective reactions and misunderstandings. All the hot-line terminals in the U.S. and the U.S.S.R. are in prime target areas. From these terminals the signals would go out over vulnerable landlines and from satellite ground stations. Two-way HF radio communication requires nothing more than that the ionosphere recover from its nuclear-explosion-induced disruption. It could therefore serve to connect the leaders of the two sides after a nuclear exchange, if appropriate agreements were made beforehand between the two countries.




**HYPOTHETICAL PROBABILITY TREE illustrates schematically the possible outcomes of a limited nuclear attack on the U.S. According to the prevailing strategic doctrine, an "appropriate" U.S. counterattack would induce the U.S.S.R. to end hostilities without triggering an all-out nuclear exchange. In a real war, however, there would be a chain of individual events involving command, control, communications and intelligence ($C^3I$), any one of which could divert this simple scenario into other outcomes. None of the unexpected outcomes "makes sense" in terms of strategic theory, and yet taken together they are actually somewhat more likely than the predicted "coherent" outcome. At each stage in this hypothetical model the "wrong" outcome was arbitrarily assigned a probability of 10 percent.**

In many ways the most vital $C^3I$ challenge is not war itself but the prelude to war. Effective crisis management presents a host of $C^3I$ issues. Good crisis communications within the U.S. Government and between governments is essential. "Hair triggers" built into nuclear forces put excessive demands on the $C^3I$ system in a crisis. The realization that nuclear weapons are subject to preemptive destruction (as in the case of silo-based ICBM's) or to capture (as in the case of battlefield nuclear weapons) may lead some national leaders to conclude that rapid decisions based on spotty information are necessary. In such conditions the choices might appear to be "use 'em or lose 'em."

In peacetime the very possession of nuclear weapons imposes on the nations having them the gravest demand for responsible $C^3I$. Procedures and technical systems must prevent unauthorized individuals, be they technicians or generals, from launching nuclear weapons. The U.S. military requires that all tasks involving nuclear weapons, from routine repairs to missile countdowns, be accomplished by two people rated in the same specialty. Many nuclear weapons have additional technical devices to prevent their crews from arming them without "enabling" codes from higher commanders. Nuclear-weapons systems must also be immune to accidental launching or detonation. Finally, nuclear weapons must be protected from theft by terrorists, foreign nations or even an ally who abruptly turns hostile.

If advanced technology sometimes seems to heighten the dangers of nuclear war, it seems to have a clear opportunity to help in improving strategic $C^3I$. Improvements in reliable sensing and data processing, resilient communications systems and survivable command posts deserve public support. No single component can be made absolutely invulnerable, but disruption can be made difficult, time-consuming, costly and above all uncertain to the adversary. Just as important is sensitive attention to procedures. Arms-control agreements, by limiting the deployment of "hair trigger" weapons, by enacting confidence-building measures and perhaps by limiting antisatellite systems, could also help the $C^3I$ system to do its job.

Although money and effort can and should be spent to improve $C^3I$, no level of effort can ever dispel the fundamental unpredictability of nuclear war. In most cases one cannot foresee precisely the physical effects that would determine the $C^3I$ system's behavior, much less predict the interaction of people and machines in chaotic circumstances they can never fully anticipate. Some observers go so far as to suggest that improving $C^3I$ will foster an illusion of control over nuclear war. Although this is probably too extreme a view, it is impossible to have confidence in any one theory about nuclear war, and confidence in strategic theories invariably declines as they are elaborated and detailed. In a fundamental sense there are no experts on nuclear war and no predictable details. One can only hope that the situation remains that way.

# Collapse and Formation of Stars

*Hidden from observation, this process can nonetheless be modeled on high-speed computers. Pictures that emerge yield insight into the formation of our own solar system*

by Alan P. Boss

What are the early stages in the formation of a star? What determines whether a cloud of star-forming matter will evolve into one, two or several stars? Because clouds of gas, dust and debris largely obscure all but the initial and final stages of the birth of a star, these questions have so far not been answered by direct observation. Theoretical modeling offers a way to circumvent this obstacle, although not an easy one. Each model requires the execution of more basic calculations than were performed by the entire human race before 1940. Today, run on sophisticated computers, such models reveal the various stages through which a star passes as it evolves. They also give a preliminary picture of how our own solar system formed.

Stars form when nebulas (interstellar clouds of gas and dust), or parts of nebulas, collapse. Although these clouds are too dense for optical telescopes to penetrate, the more diffuse clouds are transparent to millimeter-wavelength radiation. A telescope sensitive to millimeter radiation can therefore be used to observe nebulas in which stars are on the verge of forming. The clouds are also partially transparent to infrared light, and so observations in the infrared wavelengths can be made of newly formed stars within the parent nebulas. These observations yield the basic data with which any theory of stellar formation must reckon: the initial conditions under which stars form and the characteristics of the newly formed stars. Unfortunately there is still a difference of a factor of almost $10^{20}$ between the density of a star-forming cloud and that of the young stars that can be observed with infrared radiation; it has been impossible to date to view the cloud as it collapses through this range of densities. Consequently stars cannot be observed as they form.

Since the late 1960's, astrophysicists have developed increasingly sophisticated computer models of the events that take place between the two observable stages of stellar formation. Such models are based on systems of equations that describe the behavior of nebular gas and dust under the influence of many different forces; the solution of these equations can require roughly one million million basic operations. Even with a high-speed computer the calculation of one model can take several months.

Among the most important advances has been the use of increasingly realistic descriptions of the parent clouds. The early models pictured a spherically symmetric cloud with no rotation; at the next stage of complexity it was assumed that the cloud rotates but remains symmetric about its axis of rotation; in the most recent models the initial cloud rotates and is completely asymmetric.

These models have shown that a collapsing cloud will generally pass through two phases of rapid contraction (called dynamic collapse); phases during which outlying matter accumulates around a stable core follow each dynamic-collapse phase. In either phase of dyamic collapse the cloud might fragment into two or more protostars; whether or not the cloud fragments depends on such variables as its size and rate of rotation. It might also collapse in a way that produces a single protostar.

In fact such single stars are rare: in spite of the appearance of the night sky to the unaided eye, most stars are actually binary. (A binary system consists of two stars orbiting about each other. Often the members of binary systems are too close to be discerned without large telescopes or spectroscopic equipment.) Our sun, as a single star, is part of a minority population. The clouds that do not fragment are thus particularly interesting: they may rep-

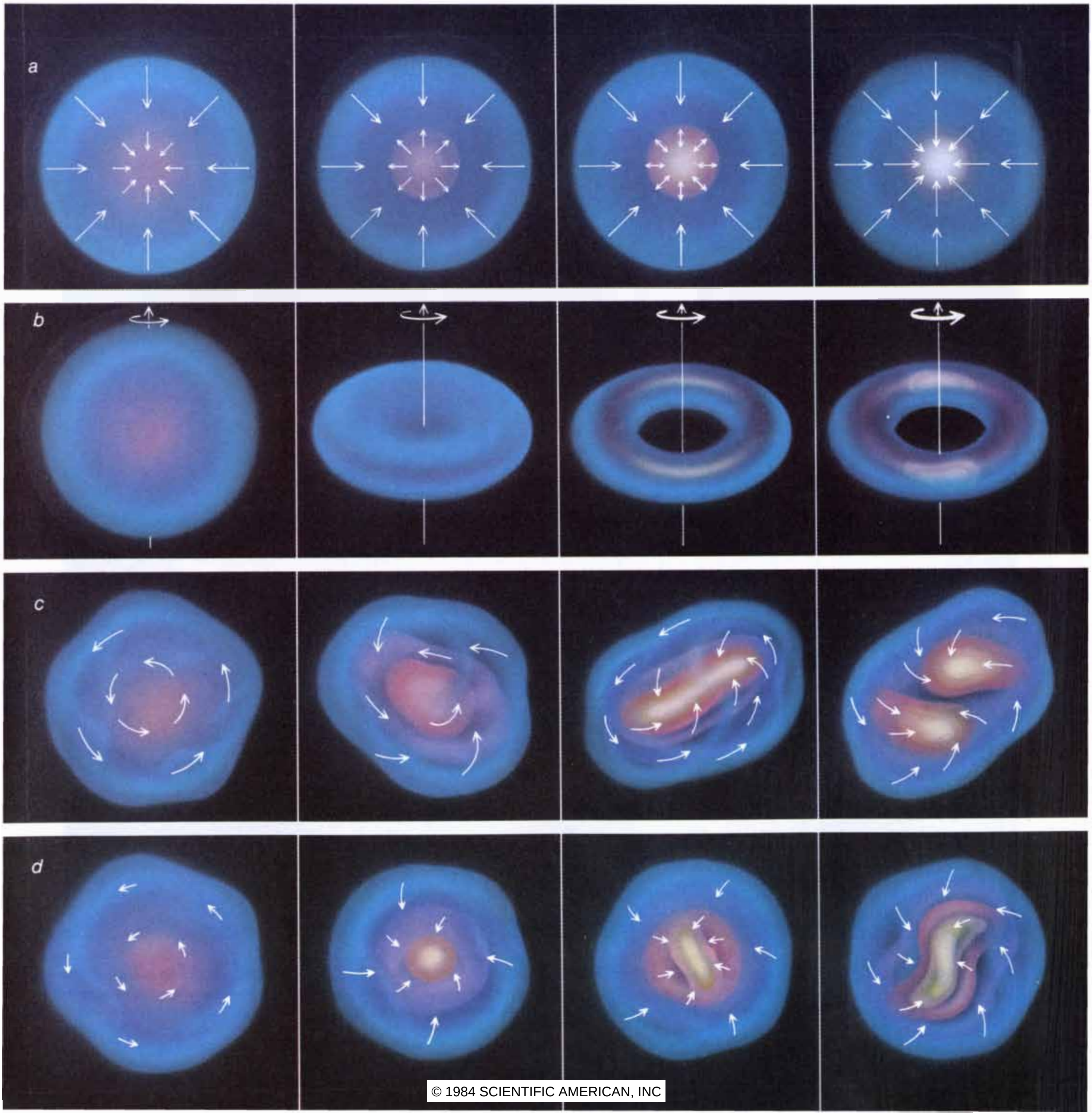
**MODELS OF STAR FORMATION grow in accuracy and detail as astrophysicists employ increasingly realistic pictures of a star's parent dust cloud. The earliest model (*A*) pictures the cloud as a perfect sphere that does not rotate. The first panel shows the dynamic-collapse phase, in which gas and dust fall rapidly toward the cloud's center. When the center becomes so dense that it is opaque to infrared radiation (*second panel*), the compressional energy produced by the collapse can no longer radiate from the cloud. Instead it adds to the thermal energy, increasing gas pressure and stopping the dynamic collapse: a first core is formed. Eventually (*third panel*) the first core becomes hot enough so that diatomic hydrogen molecules break up into single atoms. As it breaks up, the hydrogen absorbs heat and so the core's temperature falls. Pressure inside the core drops rapidly until it is no longer able to withstand the force of gravity, and a second dynamic collapse occurs. Once all the hydrogen is dissociated (*fourth panel*) the second collapse halts and the final core forms. In a more sophisticated model (*B*) a cloud that is symmetric about an axis rotates. Matter along the axis collapses faster than matter away from the axis (which feels an apparent "centrifugal force"), and so the cloud assumes a lozenge shape (*second panel*), which eventually turns into a ring (*third panel*). If the rotating cloud is even slightly irregular, the ring may break up into two or more fragments (*fourth panel*). In *C* a fully asymmetric, rotating cloud flattens out (*second panel*). The cloud becomes increasingly distorted about its rotation axis, so that it forms a bar shape (*third panel*). The bar becomes denser and more elongated as the cloud collapses, and then it breaks up into a binary protostellar system (*fourth panel*). If an asymmetric cloud rotates very slowly (*D*), it may collapse to form a single protostar. The collapse (*first, second and third panels*) is similar to that shown in *C*, but the cloud in *D* does not become as elongated and hence will not fragment; a single protostar forms. The center will rotate faster than the outer regions, and so the bar winds into a spiral; the central region may transfer angular momentum to the slower outer regions as it collapses to form a star.**

resent models for the formation of our own solar system.

How a cloud fragments is one of the two fundamental characteristics of stellar evolution that a theory must be able to describe. Interstellar clouds can be as large as 100,000 times the mass of the sun—quite massive compared with stars, which are seldom larger than about 10 times the solar mass. In addition most stars in the disk of our galaxy seem to form in clusters containing about 100 stars. These two observations suggest that interstellar clouds fragment into many protostars.

The second fundamental characteristic that must be described concerns angular momentum. In rough terms, the angular momentum of a spinning body is a measure of how much mass in the body is spinning, how fast that mass is spinning and how large the body is. According to observational evidence, interstellar clouds have up to $10^5$ times as much angular momentum per unit of mass as their progeny stars. Any theory of star formation must therefore describe how a cloud disposes of a considerable amount of angular momentum before it collapses to form a star or several stars.

One of the first sophisticated computer models of star formation was produced in 1968 by Richard B. Larson of Yale University. He developed a detailed model for the contraction of a spherically symmetric, nonrotating cloud. An important product of



his work was a picture of the so-called dynamic-collapse phase of star formation. The dynamic-collapse phase is a period of rapid contraction that can be explained by the interplay of two major forces: gravity, which tends to contract the cloud, and thermal pressure, which is the tendency of hot gas within the cloud to expand. Larson showed that the dynamic-collapse phase is due in part to the way the relation between these two forces changes because of the flow of radiation within the cloud.

The outer shell of a very diffuse dust cloud is transparent to ultraviolet radiation from neighboring stars, and hence it tends to be heated substantially by such radiation. After gravity has compressed the cloud to the density of a dark cloud, the cloud becomes opaque to ultraviolet light, eliminating this source of heating. It is still transparent to infrared radiation, however, and so dust grains in its interior are able to radiate heat energy out of the cloud in the infrared portion of the spectrum.

Thus as the density of the cloud increases, its temperature drops, down to a minimum of about 10 degrees Kelvin (degrees Celsius above absolute zero). The cloud then enters an "isothermal phase" during which the temperature remains at 10 degrees as the cloud collapses through a wide range of densities, from approximately $10^5$ to $10^{11}$ atoms per cubic centimeter. As the cloud grows smaller and denser, the gravitational force becomes stronger, eventually overwhelming the thermal pressure. The result is a dynamic collapse, in which the gas and dust fall into the center at rapidly increasing velocities. As the gas and dust fall in, the density of the cloud's center increases.

When the central regions of the cloud become dense enough to be opaque to infrared radiation, the dynamic-collapse phase ends. The collapse of the cloud has generated a great deal of heat because of the compressional work gravitational forces perform on the gas. During the isothermal phase this heat was radiated out of the cloud as infrared light; once the radiation can no longer easily escape from the cloud, the temperature and pressure begin to rise. When the center of the cloud reaches a temperature of

**YOUNG STARS** have been forming in the cluster NGC 2264 for about 20 million years. The brightest stars in this cluster occupy positions on the main sequence of stellar evolution; less luminous stars are still contracting to the higher densities and temperatures at which thermonuclear reactions begin. This optical photograph was made by David F. Malin of the Anglo-Australian Observatory.

about 100 degrees K. and a density of about $10^{14}$ atoms per cubic centimeter, thermal pressure becomes great enough to overcome the gravitational force and to stop the cloud's dynamic collapse. The region in which the dynamic collapse halts has a radius of about five astronomical units (one astronomical unit, roughly 93 million miles, is the mean distance from the earth to the sun). This region is called the first core; matter in the outer regions, still transparent to infrared radiation, continues to fall inward, accumulating at the core.

The first core is in a quasi-equilibrium state: the matter deepest within the core flows alternately inward and outward, producing a periodic increase and decrease in density.

As matter from the envelope continues to build up at the core, the center of the core becomes progressively denser and hotter. Eventually the central temperature and density become high enough so that the diatomic hydrogen molecules dissociate into single atoms of hydrogen. At this stage the temperature of the cloud is about 2,000 degrees K. and its density is roughly $10^{16}$ atoms per cubic centimeter.

Because hydrogen absorbs energy as it dissociates, the temperature of the first core drops and there is less thermal pressure to support the mass of the cloud. Consequently the first core enters a second phase of dynamic collapse. The innermost regions fall in rapidly until the core reaches densities of approximately $10^{24}$ atoms per cubic centimeter (roughly the density of water) and temperatures of about 100,000 degrees K. At this point thermal pressure again becomes sufficient to counteract the force of gravity that has been pulling matter inward. A second core, smaller than the first, is therefore able to form. Initially this core contains only a small fraction of the total cloud and is a few times the size of the sun. The remainder of the cloud, however, continues to fall inward and to enter the second core. As this falling matter accumulates, the second core replaces the first core, which disappears.

After the second core is formed and the remainder of the cloud collapses around it, the protostar enters the main sequence of stellar evolution. The entire dynamic collapse has occupied roughly 100,000 years.

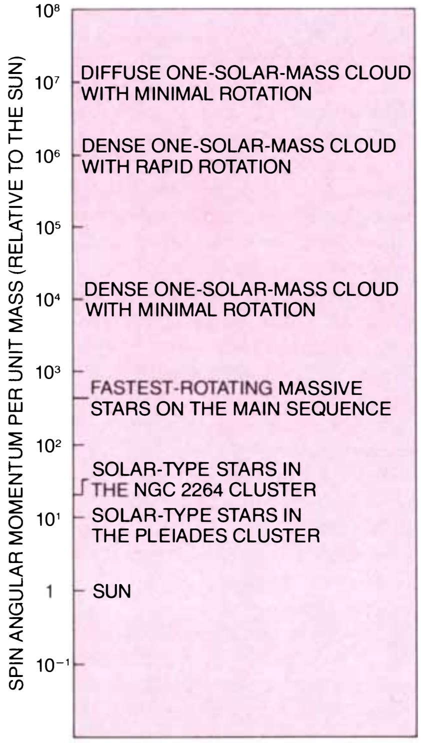
Larson's description of the collapse of spherically symmetric clouds, which I have outlined above, is basically consistent with observation. That is, it yields models of stars whose luminosities and surface temperatures fall within the observed ranges for young stars. Yet the picture of a perfectly spherical, nonrotating cloud is clearly a highly idealized one. Rotation and inhomogeneity within the cloud, neither of which is included in Larson's model, have important effects on the rate and type of collapse. It is worth noting that the nonrotational, spherically symmetric model cannot explain either fragmentation or the question of excess angular momentum.

The next step toward theoretical accuracy was taken in 1972 by Larson and in 1976 by David C. Black of the National Aeronautics and Space Administration's Ames Research Center and Peter H. Bodenheimer of the Lick Observatory of the University of California at Santa Cruz. These investigators studied the collapse of a rotating cloud; to keep the model relatively simple, they assumed that the cloud was symmetric with respect to its axis. They found that a rapidly rotating dense cloud may collapse, in several stages, to form a ring. Under certain conditions that ring may fragment into a system of several protostars.

In the first stage, matter along the axis of rotation collapses toward the center in the same way as the matter of a nonrotating cloud collapses. Matter distant from the axis collapses more slowly, because much of the gravitational force that would ordinarily pull it toward the center is needed simply to hold it in orbit about the axis of rotation. In other words, because the cloud is spinning, matter in it feels an apparent "centrifugal force": the matter would normally tend to fly off along a straight trajectory; gravity overcomes that tendency and bends the trajectory into a circle. The faster the material is moving and the smaller the orbit is, the more gravitational force is needed to maintain that orbit and the less gravitational force is available to pull the matter inward.

Since matter along the axis collapses faster than matter farther from the axis, the once spherical cloud flattens out, forming a lozenge-shaped cloud, which grows progressively flatter and more disklike as it collapses. Eventually, for reasons that were first described by Joel E. Tohline (now at Louisiana State University at Baton Rouge) and me, the disk forms a ring.

Tohline and I showed that the ring develops because of an interplay between the forces of gravity and the law of the conservation of angular momentum. The angular momentum of a spinning body depends in part on the distance between the rotating matter within that body and the axis of rotation. Since the angular momentum of an isolated spinning body must remain constant, matter that falls in toward the center must orbit faster as it falls. This means that orbiting matter cannot fall all the way to the center: as it accelerates, more of the gravitational force is necessary to keep it from flying off. Eventually the falling matter will reach "centrifugal equilibrium," where the force of gravity is precisely sufficient to maintain the matter's orbit and hold it at a constant radius.




**ANGULAR-MOMENTUM problem in the theory of star formation arises because stars have much less angular momentum than the clouds from which they form. The chart gives the angular momentum per unit of mass (in units of the sun's angular momentum) for a variety of clouds and stars.**

During the collapse of a rotating cloud some of the matter falling toward the center reaches and passes the radius of centrifugal equilibrium. Since the force of gravity is not strong enough to hold this matter in a small orbit, the matter stops collapsing inward and begins to flow away from the center (under the influence of an apparent "centrifugal force").

Meanwhile other material that is farther from the axis is still falling inward. In the resulting collision between the mass falling inward and that flowing outward a significant amount of mass accumulates away from the axial center of the cloud. If the accumulation is large enough, the gravitational force it exerts will attract the rest of the falling matter as well as the matter from the central regions. The result is a growing ring of gas and dust around an empty central region.

Thomas L. Cook of the Los Alamos National Laboratory and Michael L. Norman, now at the Max Planck Institute for Physics and Astrophysics in Munich, have shown that such a ring might eventually fragment: if the ring is not perfectly symmetric about its axis, accumulations will form along the circumference, which eventually break the ring into a system of many protostars. Cook and Norman found that the spin angular momentum of each fragment will be reduced by a factor of 10 from the spin of the initial cloud. Their model shows that the rest of the cloud's angular momentum goes into the fragments' orbital motion about one another.

The next theoretical advance took place in 1979, when it became possible to model completely asymmetric, rapidly rotating clouds. Bodenheimer, Tohline, Black and I found that some collapsing clouds will fragment without forming a ring; instead irregularities in the cloud can grow large enough to fragment it. The process would take roughly the same amount of time that an axially symmetric cloud requires to form a ring. We also found that those clouds that tend to fragment without forming rings usually evolve into binary systems rather than systems consisting of three or more members; apparently the first two accretions that form will pull the rest of the gas and dust toward themselves.

The initial fragments that form from rapidly rotating clouds typically have a mass of approximately one-tenth of the initial cloud mass and, as in the case of ring fragments, their spin angular momentum is much less per unit of mass than that of the original cloud. Furthermore, each of the fragments is likely to undergo a second dynamic collapse. As each one collapses, it breaks up into yet another set of fragments. These subfragments are themselves likely to collapse and fragment.

This hierarchy of repeated collapse and fragmentation was hypothesized by Bodenheimer in 1978, before the numerical calculations of completely asymmetric clouds were possible. Its confirmation by modeling resolves both the fragmentation and the angular-momentum questions. If a cloud undergoes a cascade of many fragmentations, it could collapse to form a modest number of protostars whose spin angular momentums would be fairly close to those of some observed rapidly rotating stars. Furthermore, the hierarchical theory suggests, as observation confirms, that far more binary than single stars will form.

As useful as they are, computer models of asymmetric clouds have until recent years suffered one major flaw. Unlike Larson's model of the perfectly symmetric cloud, these models did not take thermodynamic factors into account. That is, they have not modeled the heating and cooling of various portions of the cloud as it is controlled by the flow of electromagnetic radiation. The flow of radiation is in turn dependent on the opacity and density of gas and dust particles within the cloud, factors that change as the cloud collapses. Because of this flaw, computer models were able to consider only the isothermal phase (that part of the first dynamic-collapse phase when the temperature of the cloud remains constant), in which the effects of radiation can be neglected.

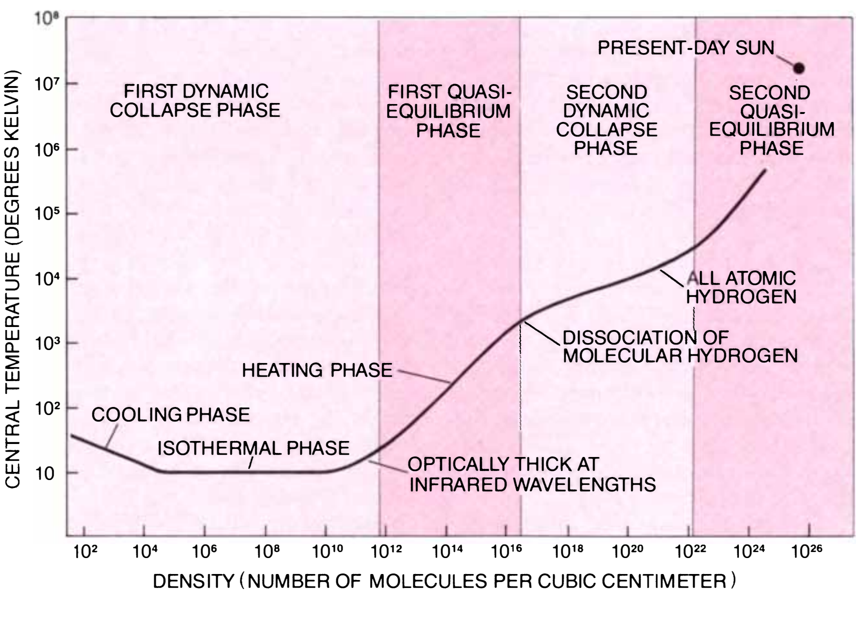
My most recent work remedies this shortcoming: it consists of a detailed analysis of the thermodynamics of asymmetric clouds. More advanced methods make it possible to follow collapsing clouds through the isothermal phase and into the next one. In this phase the opacity rises, the first core is formed and the first stage of dynamic collapse and fragmentation ends in the central region of the cloud.

My calculations have shown that there are certain types of clouds that probably will not fragment at all but instead will collapse to form single protostars. For example, a dust cloud with less than one-tenth of the mass of the sun will not undergo the hierarchy of repeated collapse and fragmentation. Likewise, a slowly rotating cloud that is somewhat more massive may collapse to form a single protostar.

This kind of cloud will flatten into a disk, which will gradually take the shape of an elongated bar. Because of the conservation of angular momentum, the inner part of the protostar will rotate more rapidly than the outer part; consequently the bar eventually elongates into a pair of spiral arms.

The inner region of the spiral protostar will transfer some of its angular momentum to the slower-spinning outer region by way of gravitational torques; that is, the gravity of the slower-spinning outer part will pull on the faster-spinning center, slowing its rate of rotation [*see illustration on opposite page*]. As the matter close to the axis of rotation slows it is able to collapse further. The protostar should then be able to contract the rest of the way to stellar densities without undergoing fragmentation due to excess angular momentum; this result has yet to be shown conclusively.

These developing models of general star formation can be applied to a specific case—the formation of our own sun and the solar system. There are three primary models of its preliminary stages. The first of these models, which suggests that the sun was origi-

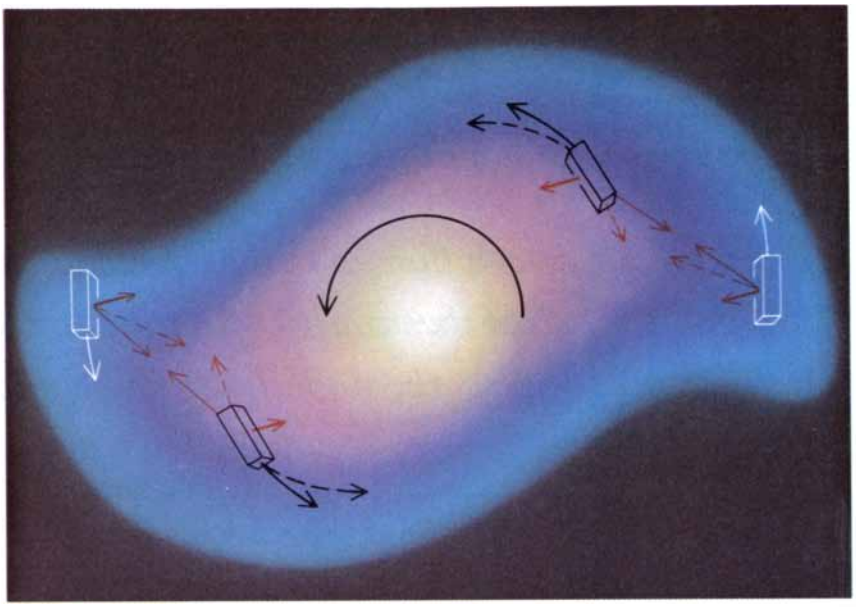



**RELATION BETWEEN TEMPERATURE AND DENSITY changes as a spherically symmetric cloud collapses. An interstellar cloud must increase in density by a factor of about $10^{24}$ and in temperature by a factor of $10^6$ before it can collapse to form a star. The conditions in the center of the present-day sun are shown for comparison (*upper right*).**

nally part of a multiple stellar system, is the least likely. The second and third models, which suggest respectively that the sun evolved from a decayed binary system and that it formed out of a slowly rotating single protostar, coincide—that is, they predict essentially the same sequence of development once the protostar is formed.

According to the first model, the sun was ejected from a system consisting of three or more equally spaced protostars. Numerical modeling has shown that multiple systems will evolve into combinations of binary protostar systems and single stars; it is therefore not inconceivable that the sun was originally part of a multiple system that underwent decay.

It is unlikely that the sun formed in this manner, however. Modeling has shown that a cloud must be rapidly rotating and relatively cold in order to form a triple protostellar system. The three protostars formed would then also have relatively large rates of rotation and low thermal energies; they would themselves collapse and fragment. A rapidly spinning star with low thermal energy could avoid fragmentation only if it is relatively small (less than one-tenth the mass of the sun). Hence, according to this model, the sun would have to have formed from a protostar with very little mass and then acquired the bulk of its mass after becoming a star, which is an unlikely history.



**GRAVITATIONAL TORQUES transfer angular momentum outward along the spiral arm of a rotating protostellar system. Inner regions (*representative volumes are outlined by black boxes*) orbit faster than outer regions of the arm (*white boxes*). Gravitational interactions (*solid colored arrows*) between the inner and the outer regions result in forces (*broken colored arrows*) that slow the inner regions' rotation and speed the rotation of the outer regions. As the rotation of the inner region slows, it falls toward the center (*broken black arrows*). If the inner matter consists of a single protostar surrounded by a spiral pattern of gas, this transfer of angular momentum may enable the central regions to contract to stellar densities.**

Apparently the solar system did not develop from a higher-order system. Perhaps it was formed in the decay of a binary system. Suppose a cloud had collapsed to form a system of two protostars spaced close together. If the binary system were somehow to transmit some of its angular momentum outward, the two stars would come closer together. If the two protostars had already reached the quasi-equilibrium phase and were no longer collapsing, the decreased separation could result in the merging of the two to form a single protostar. This protostar would closely resemble the protostar that would result from the third proposed model of solar evolution, which is the collapse of a slowly rotating cloud.

As discussed above, a cloud that is rotating extremely slowly would not fragment during the dynamic-collapse phase. Werner M. Tscharnuter of the Astronomical Institute of the University of Vienna has modeled the collapse of slowly rotating, axially symmetric clouds and shown that they will not form rings. He has therefore inferred that such clouds will not fragment. My own three-dimensional calculations support Tscharnuter's results. They also show that even slowly rotating clouds will collapse into protostars with the barlike shape characteristic of the faster-rotating clouds; this means that the central portions of the protostar could transmit some of their angular momentum outward by gravitational torques, enabling the protostar to contract to stellar densities. The implication is that slowly rotating clouds will form single stars. Since single stars are in the minority, it appears that slowly rotating clouds must also be relatively rare.

Both the model of a decayed binary and that of a slowly rotating cloud converge into one model on the formation of a single protostar surrounded by a cloud of gas and dust. At this stage the protostar must still contract through an increase in density by a factor of $10^{10}$ and undergo a second collapse before it reaches the main sequence. This phase of stellar evolution has not yet been calculated rigorously with a three-dimensional model; also certain physical properties that I have not discussed here, such as turbulence and magnetic fields, may have important effects.

At the same time that the sun is forming at the center of the nebula, the dust in the outer regions will form a flattened layer and begin the process of accumulation into a planetary system. This surrounding gas and dust may be essential for forming a single star, because it provides the protostar with a way to disperse some of the angular momentum that would otherwise impede its collapse. The formation of a planetary system may thus be a natural consequence of the formation of a single star. The recent exciting discovery of a flattened layer of dust surrounding the star Beta Pictoris seems to confirm this general picture of star and planet formation.

As astrophysicists have made increasingly realistic assumptions about the dust clouds, a clearer image of the process of stellar formation has developed. The next stage in my own research will be an attempt to extend the thermodynamic models of asymmetric clouds. Until now the model has been applied only to the first quasi-equilibrium phase, the forming of the first core; next I shall examine the second dynamic-collapse phase. I believe that no further fragmentation occurs after the first core has formed, and that the protostar contracts to stellar densities, but the definite answer can come only after rigorous modeling, which may take several more years.

# The Perception of Speech in Early Infancy

*In perceiving speech human beings detect discrete phonemic categories and ignore much of the acoustic variation in the speech signal. Research with infants suggests the underlying perceptual mechanisms are innate*

by Peter D. Eimas

How is it that a child swiftly and seemingly without much effort learns to speak and understand? The process of language acquisition begins well before the first birthday, and most children use language with considerable skill by their third year. In contrast to the learning of reading or arithmetic, a child masters language without formal teaching; indeed, much of the learning takes place within a fairly limited linguistic environment, which does not specify precisely the rules governing competent language use.

A possible explanation for the swift growth of a child's language skills is that language is not as complex as is generally thought, and consequently that such simple psychological principles as conditioning and generalization account for the speed with which it is learned. But research during the past several decades on the nature of language and the processes by which it is produced and understood has revealed not underlying simplicity but increasing complexity.

Experiments carried out by my colleagues and me at Brown University and by other investigators elsewhere have supported a different explanation, one derived from the view, of which the linguist Noam Chomsky is the most notable exponent, that inborn knowledge and capacities underlie the use of language. In studies of speech perception by infants we have found these young subjects are richly endowed with innate perceptual mechanisms, well adapted to the characteristics of human language, that prepare them for the linguistic world they will encounter.

The search for inborn mechanisms of speech perception developed from studies of the relation of the speech signal to phonemes, the perceptual units that correspond to the consonants and vowels of language. Phonemes are the smallest units of speech that affect meaning: only one phonemic difference sets apart the words *late* and *rate,* yet they are entirely distinct in meaning.

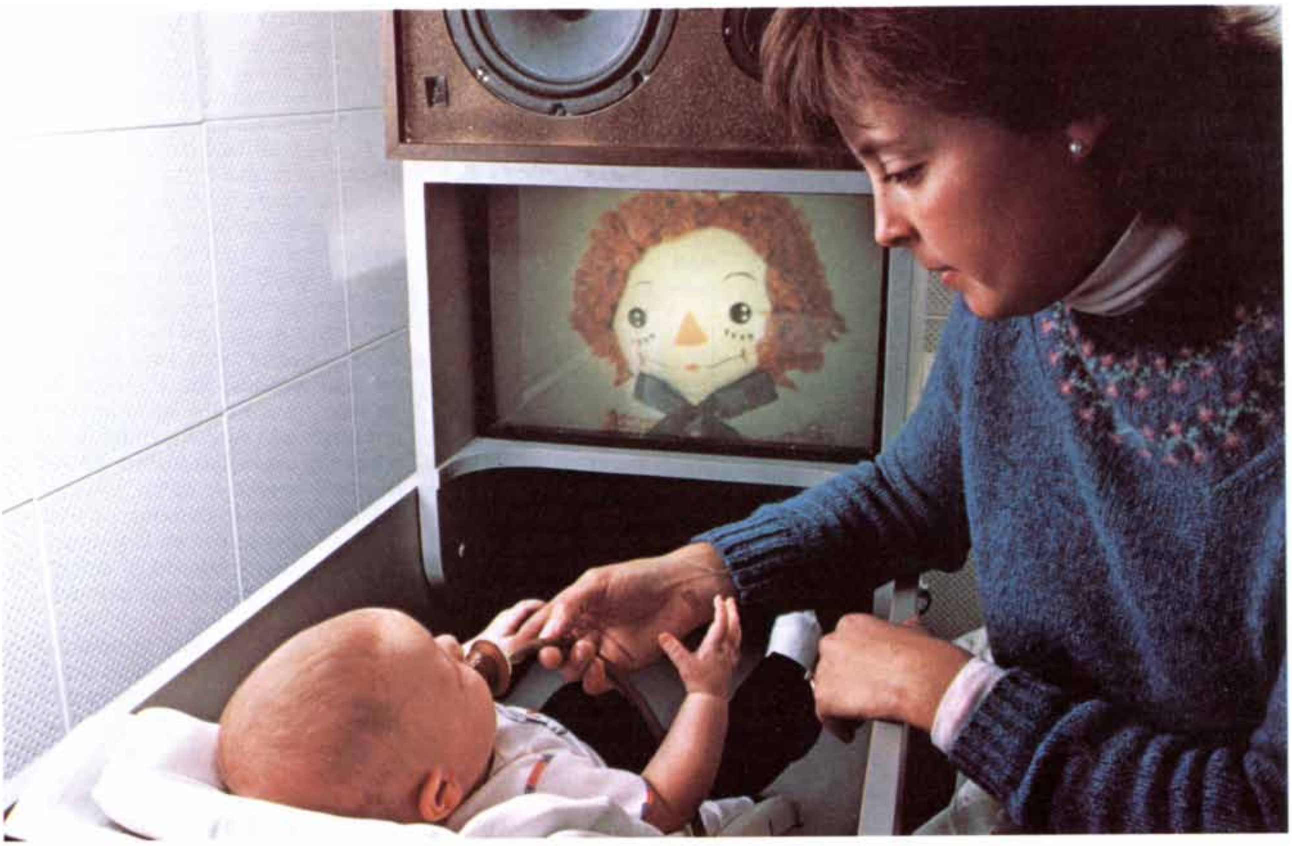
Workers at the Haskins Laboratories in New Haven, the Massachusetts Institute of Technology, Sweden's Royal Institute of Technology and elsewhere have shown that the speech signal is a complex of acoustic units: brief segments bounded by momentary pauses or peaks in intensity. The segments vary in duration and in the frequency, temporal relations and intensity of their constituent bands of concentrated acoustic energy, known as formants, and of noiselike acoustic components known as aspiration and frication. The variation in these acoustic parameters provides the information that is critical to the perception of phonemes.

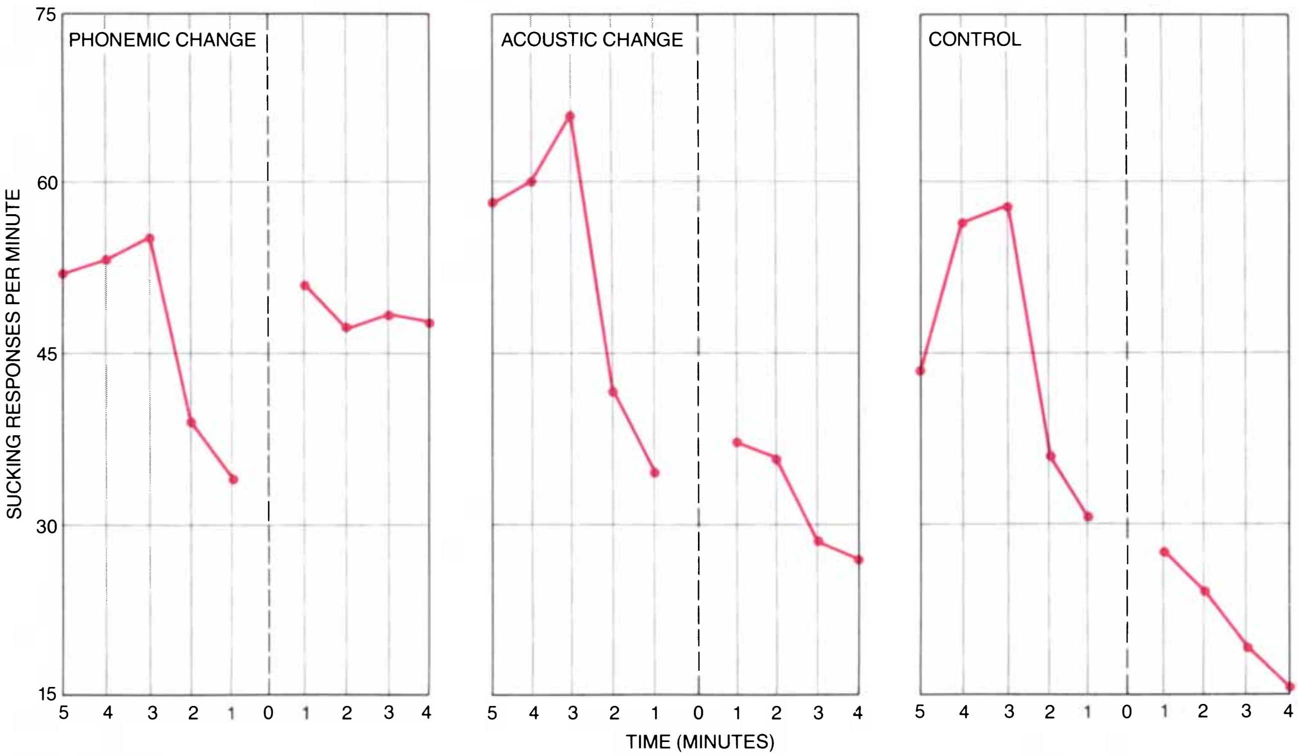
No direct, one-to-one correspondence holds, however, between individual acoustic segments and the phonemes we perceive. A single acoustic segment may encompass a consonant and a vowel; conversely, two distinct acoustic segments may contribute to a single consonantal sound. Furthermore, there is no direct relation between the segments' frequency and temporal characteristics and the phonemes we hear. A listener may recognize a range of stimuli, varying widely in a number of acoustic traits, as instances of the same phoneme. On the other hand, a small change in a single acoustic cue may in some situations change the phoneme that is perceived.

Consider the acoustic information that is sufficient to signal the distinction between the voiced stop consonant that begins the word *bin* and the voiceless stop consonant in *pin.* In both cases the speaker completely blocks the flow of air through the vocal tract just before the release of the utterance; in *bin,* however, the vocal cords begin to vibrate almost simultaneously with the release of air, whereas in *pin* vocal-cord vibration is delayed. The interval between the release of air and the onset of vocal-cord vibration, or voicing, is known as voice-onset time; it holds the crucial acoustic information that enables a hearer to distinguish *bin* from *pin.* No single value of voice-onset time defines each phoneme, however. Instead hearers typically perceive a range of values, reflecting different speakers, different instances of speech and differences in the surrounding phonemic environment, as examples of the same phoneme.

The acoustic variables that define other phonemes are similarly fluid. For example, many phonemes are differentiated by place of articulation, the site of the constriction of the vocal tract that occurs as the sound is formed; the initial sounds of *bin* and *din* are examples. Among the acoustic cues that correspond to place of articulation and enable a hearer to distinguish such phonemes are the initial frequencies of the second and third formants: the formants that fall second and third from the bottom on a scale of frequency. Again no single value of these acoustic parameters characterizes each phoneme; a range of onset frequencies can signal the same place of articulation. Yet in spite of the variation in the sounds corresponding to each phoneme we have little trouble deciding whether a speaker said *din* or *bin.* We are able in effect to listen through the variation in the signal and make categorical judgments of phonemic quality.

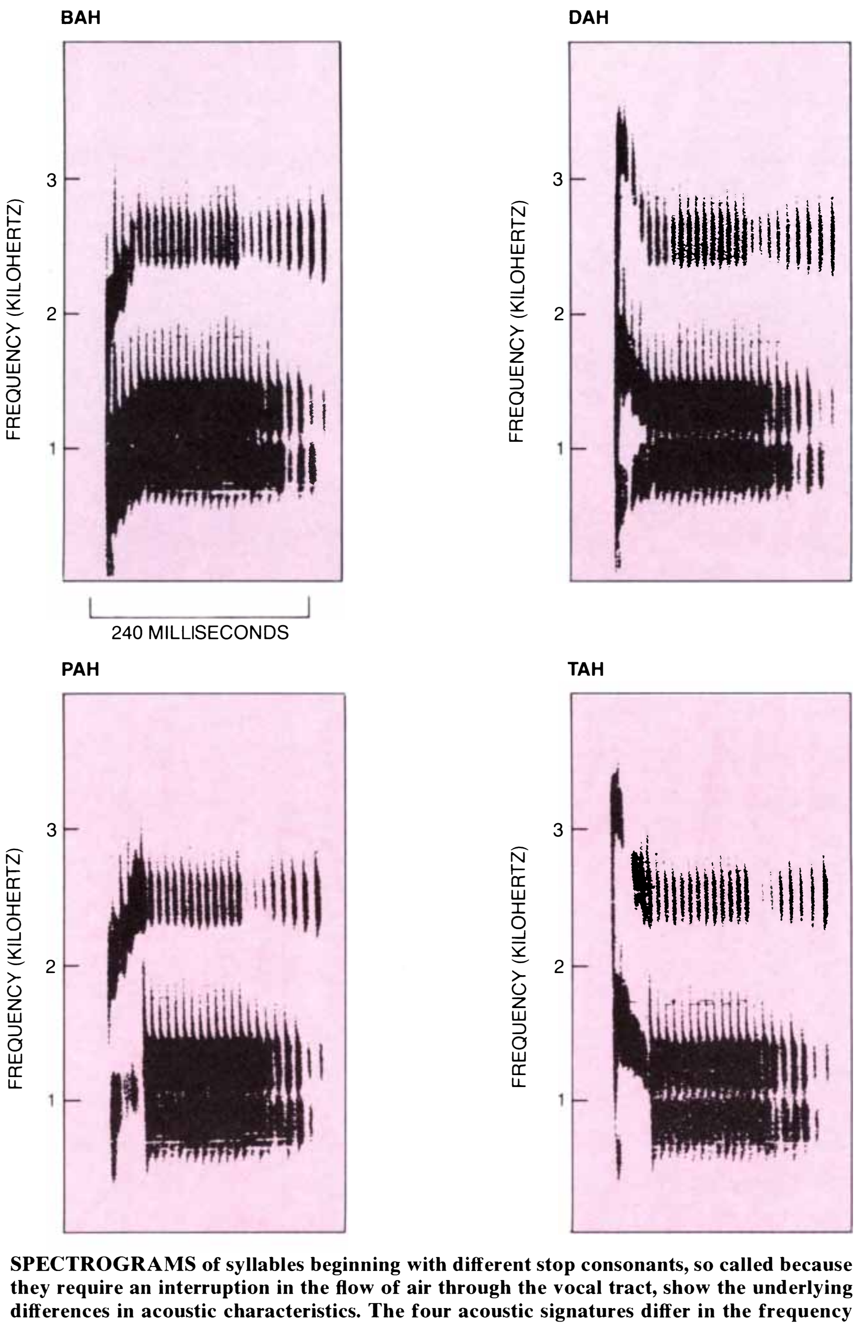
Experimental results confirm that in






**INFANTS' SUCKING RATE** indicates their response to a series of speech sounds. In the author's experimental setup (*top*) syllables of synthetic speech were played through the loudspeaker above the screen display of Raggedy Ann while a four-month-old infant sucked on a pacifier connected to recording instruments. Graphs of mean sucking rate (*bottom*), recorded under various experimental conditions with a number of infants, show that when a syllable beginning with a particular consonant was repeated, sucking rate first increased and then decreased as the stimulus became familiar. In some cases the sound changed at a time indicated by the broken line. In one group (*bottom left*) the new sound represented a different consonant; sucking rate increased sharply, showing that the infants perceived a contrast. In a second group (*bottom middle*) the stimulus differed acoustically from the preceding sound but corresponded to the same consonant, and there was little change in sucking rate. A control group of infants (*bottom right*) experienced no change in stimulus.

the perception of speech we are ordinarily aware of discrete phonemic categories rather than of the continuous variation in each acoustic parameter: we perceive speech categorically. In experiments conducted by Leigh Lisker and Arthur S. Abramson of the Haskins Laboratories adults heard computer-generated speech sounds that embodied a range of different values of voice-onset time. In spite of the many variants of voice-onset time the subjects heard nearly all the stimuli either as a voiced phoneme such as the initial consonant of BAH or as a voiceless phoneme such as the consonant that begins PAH. The boundary—the voice-onset time at which listeners began to hear PAH instead of BAH—was situated at about 30 milliseconds following the initial release of air.




**SPECTROGRAMS** **of syllables beginning with different stop consonants, so called because they require an interruption in the flow of air through the vocal tract, show the underlying differences in acoustic characteristics. The four acoustic signatures differ in the frequency and timing of their component bands of acoustic energy, known as formants. The consonants paired horizontally are distinguished by the frequency at which the formants begin, a reflection of the point within the vocal tract at which constriction occurs. The frequency of the highest formant of the sound BAH, for example, begins at about two kilohertz and then rises, while that of the third formant of DAH begins at about three kilohertz and falls. Consonants paired vertically differ in voice-onset time, a measure of the delay between the release of air and the vibration of the vocal cords. In spectrograms for BAH and DAH a voice-onset time of zero is evident in the presence of periodicity, a series of spiky vertical striations that indicate vocal-cord vibration, at the beginning of all three formants. In spectrograms for PAH and TAH there is a gap before the lowest formant appears and periodicity begins in the two higher formants, reflecting a longer delay in onset of voicing.**

To confirm the categorical nature of speech perception the experimenters asked subjects to distinguish pairs of stimuli differing in voice-onset time. If both sounds represented voicing delays of less than 30 milliseconds, the listeners generally heard them as two identical instances of BAH; if the voice-onset times of both were longer than 30 milliseconds, the listeners tended to hear two PAH's, indistinguishable although acoustically different. Only when the stimuli straddled the 30-millisecond boundary could subjects distinguish them consistently. Catherine G. Wolf, then at Brown University, obtained similar evidence of categorical perception in school-age children.

How much of this mechanism of categorical perception, which enables us to perceive speech reliably in spite of the lack of precision of the speech signal, is innate? The fact that speakers of different languages are attuned to somewhat different phonemic distinctions suggests that the influence of the linguistic environment on speech perception is powerful. Japanese speakers fail to perceive the contrast between the phonemes /r/ and /l/, a standard distinction in English; English speakers do not notice a fundamental contrast in voicing that distinguishes certain phonemes in Thai. Yet certain phonemic distinctions are present in languages throughout the world. It seemed possible to my colleagues and me that strong biological determinants, modified by later linguistic experience, might underlie the categorical perception of speech. To find out whether this is the case we did experiments with infants not yet able to speak, in whom one would expect the influence of their parents' language to be minimal.

In 1971 Einar R. Siqueland, Peter W. Jusczyk, James Vigorito and I examined the perception of voice-onset time in one- and four-month-old infants. We exposed the infants to three different pairs of sounds. The voice-onset times of one pair were 20 and 40 milliseconds; thus the stimuli fell on opposite sides of the category boundary recognized by adult speakers of English and other languages. To adult ears the stimuli sounded like the syllables BAH and PAH. In each of the other pairs, with voice-onset times of zero and 20 milliseconds and 60 and 80 milliseconds, both stimuli fell on the same side of the voiced/voiceless boundary; both were instances of BAH or PAH.

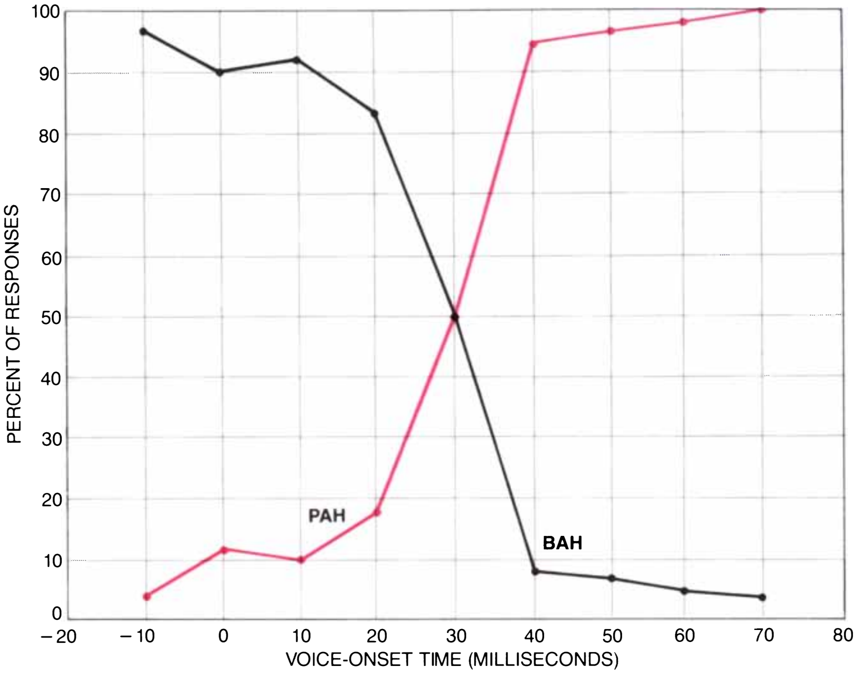
Infants a few months old cannot report their perceptions directly. In order to gauge the infants' responses to the stimuli we resorted to a methodology called the high-amplitude sucking procedure. Each infant sucked on a pacifier wired to a pressure transducer, which in turn was connected to recording instruments. We adjusted the setup's sensitivity separately for each infant so that in every case the apparatus recorded a base-line rate of sucking of 20 to 40 times a minute.

Once the experiment was under way, each time the apparatus recorded an instance of sucking one sound of a stimulus pair was played. When an infant encounters a new stimulus, its rate of sucking typically increases for several minutes, then gradually settles back to the base-line rate, presumably as a result of familiarization. When the sucking rates of our subjects fell to a preset level as they grew accustomed to the first sound, we shifted the stimulus to the other sound of the stimulus pair. If an infant grows familiar with one stimulus and then encounters a stimulus it perceives as different, its rate of sucking ordinarily increases.

The results showed that infants, like people who command a language, perceive differences in voice-onset time categorically. When both the sounds to which an infant subject was exposed lay on the same side of the 30-millisecond boundary, the shift from one sound to another evoked no increase in sucking rate. The infants appeared not to notice the change in voice-onset time. On the other hand, when the stimuli fell on opposite sides of the boundary, a sharp increase in sucking rate occurred at the shift, indicating that the infants perceived a change.

Other investigators and I have discovered further perceptual boundaries in infants' responses to the acoustic information in speech. Like adults, they respond categorically to changes in the onset frequency of the second and third formants, an acoustic cue that indicates differences in the place of articulation of a consonant. The same pattern holds in their responses to the acoustic cues that signal the distinctions between nasal and stop consonants, exemplified by the initial sounds of MAH and BAH, and between stop consonants and semivowels such as the initial sound of WAH.

It is difficult to see how learning could account for the mode of perception we have demonstrated in infants. What events during the first few weeks of life would train an infant to respond categorically to gradations of acoustic properties? A simpler view is that categorization occurs because a child is born with perceptual mechanisms that are tuned to the properties of speech. These mechanisms yield the forerunners of the phonemic categories that later will enable the child unthinkingly to convert the variable signal of speech into a series of phonemes and thence into words and meanings.




**CATEGORICAL PERCEPTION is reflected in curves showing the relative proportions of responses when children were asked to identify a synthetic speech sound with a particular voice-onset time as an instance of a voiced (BAH) or a voiceless (PAH) consonantal sound. Instead of a linear change in the percentages the curves show that at voice-onset times of less than 30 milliseconds the children almost always identified the stimulus as BAH; when voice-onset time exceeded 30 milliseconds, they tended to hear the sound as PAH. The perceptual tendency shifted abruptly at 30 milliseconds. The study, done by Catherine G. Wolf at Brown University, suggests that perceptual categories, rather than continuous gradations in the acoustic properties of the speech signal, shape the perception of speech.**

If these perceptual mechanisms do represent a biological endowment, they should be universal. The same perceptual patterns should occur in infants of every linguistic background. In research reported in 1975 Robert E. Lasky, Ann Syrdal-Lasky and Robert E. Klein, then at the Institute of Nutrition in Panama, investigated the perception of voice-onset time by Guatemalan infants, born into a Spanish-speaking environment. The group's experimental methods differed from those used in our 1971 study: in place of changes in sucking rate they used changes in heart rate as the gauge of infants' response to the speech patterns. The study also tested sensitivity to a voicing category we had omitted, one found among stop consonants at the beginning of syllables in Thai and in a number of other languages although not in English. In this so-called prevoiced category the vocal cords begin to vibrate up to 100 milliseconds before the release of air, in a kind of prefatory hum.

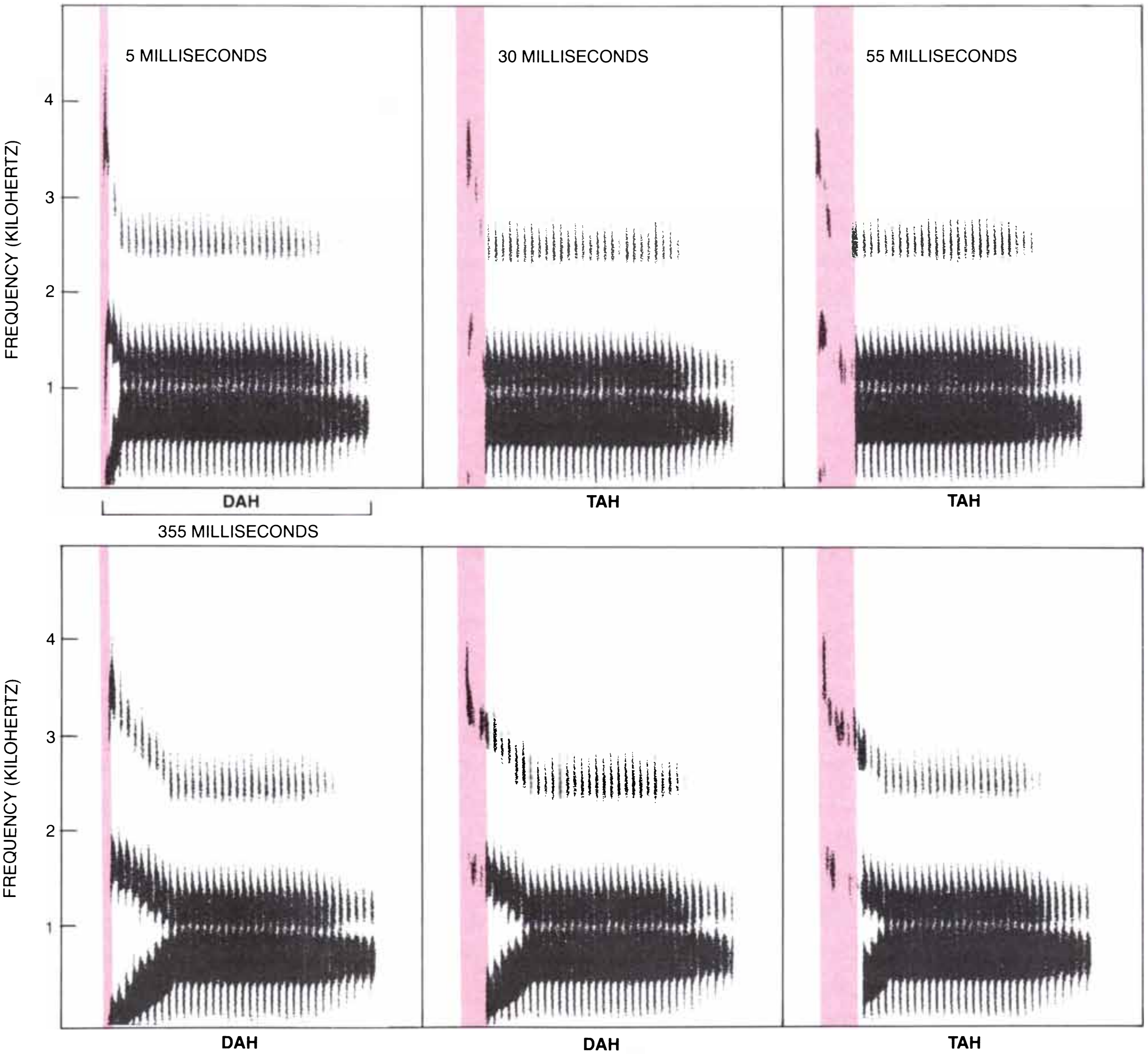
Lasky and his co-workers exposed the infants to three pairs of stimuli. In the first pair the voice-onset times fell at 20 and 60 milliseconds after consonantal release; thus the two sounds lay on opposite sides of the voiced/voiceless boundary recognized by speakers of English and other languages, although as it happens not by Spanish speakers. The stimuli in the second pair had voice-onset times of 60 and 20 milliseconds prior to consonantal release and fell on opposite sides of the prevoiced/voiced boundary of Thai. In the sounds of the final pair voicing began 20 milliseconds before and 20 milliseconds after consonantal release. Spanish speakers, in contrast to speakers of many other languages, perceive the voiced/voiceless boundary as falling between those two values.

The tracings of heart rate recorded any increases that occurred when the infants, having grown accustomed to the first sound of a stimulus pair, heard the second sound. The data showed the young subjects responded to the prevoiced/voiced distinction, with a boundary between 60 and 20 milliseconds before consonantal release, and also to the voiced/voiceless distinction with a boundary between 20 and 60 milliseconds following release. The voicing distinction peculiar to Spanish evoked no change in heart rate.

In 1976 Lynn A. Streeter, then at the Bell Laboratories, published evidence that infants born into a Kikuyu-speaking culture in Kenya display much the same perceptual pattern as the Guatemalan babies. Richard N. Aslin, David B. Pisoni, Beth L. Hennessy and Alan J. Perey of Indiana University recently completed the study of voice-onset-time sensitivity among infants in English-speaking communities by showing that they respond to the prevoiced/voiced contrast just as they do to the voiced/voiceless distinction. It appears that infants the world over are equipped with an inborn sensitivity to these three categories of voicing, whether or not the distinctions are important in their parent language.

The perception of speech is a complex and subtle process, which the studies of categorical perception described so far probe only in the simplest terms. The acoustic information that enables a listener to perceive distinctions in voicing illustrates the point. So far we have treated the essential information as a single continuum of time measuring the interval between consonantal release and the beginning of voicing. In ordinary speech, however, an interplay of temporal and spectral factors governs the perception of voicing distinctions. These acoustic properties interact in what might be called perceptual trading relations: a change in the value of one property alters the value of another at which the perceptual boundary falls.

For example, because of functional characteristics of the mechanisms of articulation, the frequency of the first, or lowest, formant rises as voice-onset time increases. Our perceptual system seems to be attuned to the relation: a frequency change can substitute for a




**SHIFT IN A PERCEPTUAL BOUNDARY can occur when two acoustic cues are altered independently. The consonants beginning the six syllables shown in spectrographic form vary in voice-onset time and in the onset frequency of the lowest formant; to adult ears the sounds are the syllables DAH and TAH. At a high onset frequency (*top row, visible in second and third spectrograms*) infants detected the DAH/TAH contrast between sounds having voice-onset times of five and 30 milliseconds. When the onset frequency was low (*second and third spectrograms in bottom row*), voice-onset time had to increase to between 30 and 55 milliseconds before the infants reacted to the phonemic contrast. Such interactions between two acoustic variables are known as perceptual trading relations.**

change in the temporal cue. When the first formant begins at a higher frequency, the effect is the same as if the voice-onset time had lengthened. As a result, at higher onset frequencies adults perceive the voiced/voiceless boundary earlier in the continuum of voice-onset times.

The same subtleties are apparent in the perceptual systems of infants. In 1983 Joanne L. Miller of Northeastern University and I showed that one perceptual trading relation found in adults also holds in infants' responses. We found that the voice-onset time at which three- and four-month-old infants recognize a shift from the voiced initial sound of the syllable DAH to the voiceless sound of TAH varies with the onset frequency of the first formant.

A second complication in the perceptual process arises from the fact that the category boundaries perceived by adults shift not only as a result of the interplay of multiple cues but also with variations in acoustic context. In this respect as well infants display the forerunners of more mature patterns of perception. Miller and I have shown that infants, like adults, distinguish the stop consonant of BAH and the semivowel of WAH differently depending on the duration of the vowel sound that follows. The acoustic basis of the distinction is the length of the formant transitions: the periods needed for the central frequencies of the formants to reach the values appropriate for the vowel that follows. In the case of BAH the formant transitions are swift; with WAH they are slower. The longer the vowel duration is, however, the slower the formant transitions must be before infants recognize a change in stimulus from BAH to WAH.

Other quite complex effects of context on the categorization of speech by infants have been demonstrated. Jusczyk and his associates at the University of Oregon found a shift in the formant-onset frequencies at which infants detect a distinction between phonemes differing in place of articulation. The boundary value varied depending on whether an additional band of noiselike acoustic energy was present, signaling a fricative rather than a stop consonant.


**BABY RECOGNIZES A PHONEMIC CONTRAST in an experiment devised by Patricia K. Kuhl of the University of Washington to investigate infants' ability to distinguish contrasting phonemes from acoustically varied instances of the same phoneme. In this case the baby, its attention held by a toy, ignored variations in speaker and intonation among repetitions of the vowel sound /a/, as in POP (*top*). When instances of the vowel /i/, as in PEEP, interrupted the sequence, the infant turned away from the toy and toward the loudspeaker (*bottom*), indicating recognition of the linguistically important contrast. The sight of a mechanical stuffed rabbit, illuminated in its case on top of the loudspeaker when the contrasting phoneme was played, served to reward the infant's accurate response.**

The complex mechanism of categorical perception enables an individual to recognize phonemes consistently in spite of great variation in crucial acoustic parameters. Other kinds of variability blur the definition of the speech signal even further. The length of syllables, along with other temporal characteristics of speech, changes with rate of speech and patterns of emphasis; wide variations in the fundamental frequency of voicing and in the spacing of resonant frequencies occur as a result of the speaker's sex, age and emotional state. Some mechanism must enable us to listen through the variation to hear the same phoneme each time it is spoken. This phenomenon of perceptual constancy cannot be investigated directly in infants. But studies of infants' ability to form equivalence classes—groups of stimuli that evoke the same response in spite of obvious differences—suggest infants possess at least the forerunners of perceptual constancy.

Patricia K. Kuhl and her colleagues at the University of Washington have investigated the formation of equivalence classes for the sounds of speech in six-month-old infants. In the first stage of each experiment the Kuhl group trained infants to turn their head 90 degrees toward a loudspeaker whenever a series of contrasting stimuli interrupted a background sound; the sight of a colorful, moving toy that appeared above the loudspeaker as the contrasting sequence was played rewarded successful responses. In one experiment identical instances of the vowel sound /a/, as in POP, served as the background stimulus; identical versions of /i/, as in PEEP, provided the contrast. Once the training was complete the stimuli were varied: the vowels, /a/ and /i/, remained the same but the infants now heard both vowels in a variety of voices and intonations. Sequences without contrasting stimuli, in which every sound was a variant of /a/, served as controls.

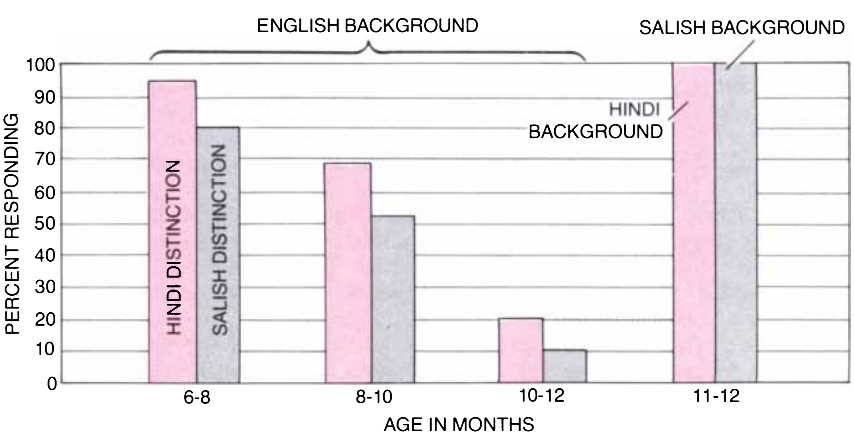
The infants' success in singling out

contrasting stimuli and ignoring within-category acoustic variations during the control trials was impressive. When both inappropriate head-turnings and missed contrasts were counted, they averaged about 80 percent correct; in seven out of eight cases the infants scored better than if their responses had reflected chance. When Kuhl and her colleagues repeated the experiment with the acoustically less distinctive vowels /a/ and /ɔ/ (as in PAW), the infants still could detect equivalent sounds, although less reliably; the proportion of correct responses fell to 67 percent and only four out of eight infants bettered the expected score for random responses.

When both the background and the contrasting sequences included arbitrarily chosen variants of both /a/ and /i/, however, the infants could not learn to differentiate members of the two sequences, in spite of the reward elicited by a correct response. They could not be trained to recognize an arbitrary grouping of sounds that had no linguistic property in common. They could respond correctly, indicating they had organized diverse stimuli into equivalence classes, only when the background and contrasting sequences represented different categories of speech. The finding is further evidence that long before infants can speak and understand they are particularly sensitive to the acoustic distinctions crucial to the comprehension of speech. It adds weight to the case for a set of inborn mechanisms that are specialized for speech perception.

The diversity among the sound systems of human languages makes it clear that environmental factors affect the perceptual dispositions with which an infant is born. What happens as the linguistic environment created by a child's parents and companions interacts with inborn perceptual mechanisms? It appears that perceptual horizons narrow as a child learns his or her native language. The child retains and probably sharpens those perceptual capacities that correspond to phonemic distinctions in the parental language but loses the ability to detect distinctions that do not occur in the native language.

Studies of voice-onset-time perception testify to the decline in some discriminative powers as the infant develops. While infants from diverse linguistic backgrounds respond to contrasts in prevoiced, voiced and voiceless initial consonants, adult speakers of some languages, including English, recognize only the distinction between the voiced and the voiceless categories. Although native adult Japanese speakers are virtually unable to perceive the distinction between the sounds of /r/ and /l/ without special training, I have found that the distinction is among those to which American infants—and presumably Japanese infants—have an innate sensitivity. Similarly, research by Janet F. Werker of Dalhousie University in Nova Scotia and Richard C. Tees of the University of British Columbia showed that six- to eight-month-old infants from an English-speaking background readily distinguish phonemic contrasts in Hindi and Salish, a North American Indian language. When they were tested again at the age of 12 months, the same infants, like English-speaking adults, did not detect the contrasts to which they had earlier been sensitive.




**WANING OF UNUSED PERCEPTUAL POWERS is evident in the responses of infants from an English-speaking background to linguistic contrasts that are foreign to English. When Janet F. Werker of Dalhousie University in Nova Scotia and Richard C. Tees of the University of British Columbia simultaneously tested infants in different age groups, the proportion responding to consonantal contrasts from Hindi and Salish, a North American Indian language, fell rapidly with age. One-year-old Hindi and Salish infants, in contrast, retain the capacity to perceive the linguistic contrasts native to their respective languages.**

The decline in perceptual abilities through exposure to a restricted environment is familiar. When kittens are raised wearing goggles that limit the visual input of one eye to a series of horizontal stripes and that of the other eye to vertical stripes, corresponding areas of the visual cortex lose their sensitivity to stripes running in other directions. Such losses seem to be irreversible, no matter how varied the animal's later surroundings. In contrast, we can recover at least some of our initial capacities to detect the acoustic information underlying phonemic contrasts. For instance, when the acoustic information critical to phonemic distinctions in Hindi and Salish is embodied in sounds that usually are not heard as speech, English speakers can detect differences to which they are ordinarily insensitive.

Apparently the restricted linguistic environment of one's native language does not inactivate unused perceptual mechanisms completely. We learn to listen primarily for the acoustic distinctions that correspond to phonemic contrasts in our own language. Given the right task or instructions, however, we can detect unfamiliar acoustic distinctions even though we do not perceive them as marking phonemic contrasts. Furthermore, with enough experience the perception of non-native distinctions begins to operate at the phonemic level: after considerable experience with spoken English, native speakers of Japanese can distinguish the phonemes /r/ and /l/ categorically and almost as accurately as native English speakers. The fact that perceptual mechanisms available to us as infants can still operate in adulthood, after long disuse, confounds hypotheses that early experience with language immutably alters some of the mechanisms of speech perception.

The most dramatic demonstration of the innate mechanisms of perception other workers and I have studied, however, takes place in infancy, as a child begins to learn its parents' language. It is now clear that an infant is born with many of the underpinnings of later speech perception and comprehension. It may be that like the specialized anatomy of the vocal tract and the speech centers in the brain these innate perceptual capacities evolved specifically for the perception and comprehension of speech. They are an evolutionary answer to the need for each infant to acquire its parents' language and culture as early in life as possible. The effectiveness of these mechanisms is reflected in the swiftness with which a child joins the community of language.


BREAKTHROUGH:
ADD LIFT TO LET
A HUGE PLANE LAND
ON A POSTAGE STAMP.
Since most large planes need a lot of room to touch down and come to a stop, they can't always land where they want to.
But we've developed a way for very big planes to land on very short runways.
The plane's engine exhaust is blown across special titanium wing flaps to create added lift at low speeds. And that allows large airplanes to approach at steeper angles and make more precise touchdowns.
This revolutionary technology, seen on this aerodynamic test model, will let the McDonnell Douglas C-17 deliver large equipment virtually anywhere. It will be able to fly from the United States directly to more than 10,000 runways around the world.
We're creating breakthroughs not only in aviation but also in health care, automated factories and information systems.
We're McDonnell Douglas.
USA
20c
MCDONNELL
DOUGLAS

# SCIENCE AND THE CITIZEN

## *Goal to Go*

This month will see the second meeting of a National Academy of Sciences (NAS) study panel commissioned to advise the National Aeronautics and Space Administration (NASA) on "Major Directions for Space Sciences: 1995–2015." The meeting comes as NASA faces sharp criticism.

One of NASA's most persistent critics has been Freeman J. Dyson of the Institute for Advanced Study in Princeton. Dyson's main charge is that space-science projects have become large and inflexible. Such projects take so long to complete that "in most of the major missions," Dyson says, "the instruments were designed to answer questions that seemed important ... a decade or more earlier.... Space science begins to look like a two-horse shay with a cart horse and a racehorse harnessed together."

As evidence that NASA plans too far in advance, Dyson points to the title of the NAS study. The panel is considering only those proposals that could be undertaken after 1995; any missions to be launched before then are already in the planning stages. Dyson would rather see space science move in small steps, allowing each simple, specialized mission to determine the questions the next project should answer.

Although few investigators would like to see large projects abandoned, there are many who believe small projects deserve increased emphasis. Thomas M. Donahue of the University of Michigan, who chairs the NAS Space Science Board, points out that some important projects, such as the Space Telescope, could not be reproduced on a smaller scale. He adds, however, that many simple missions that could be very useful are being ignored: "Big missions are wonderful, but they are not all we need for a balanced, healthy space program."

Riccardo Giacconi of Johns Hopkins University, director of the Space Telescope Science Institute, agrees there are inherent dangers in large projects, but he says some scientific questions cannot be answered in any other way. As examples he mentions the particle accelerators that are needed in high-energy physics. Giacconi believes astronomy, like physics, has reached the point where it must occasionally rely on expensive apparatus.

James M. Beggs, NASA's chief administrator, says the argument that some space-science projects are too large is "a legitimate criticism of the program, but as our knowledge base increases ... the equipment we need grows more sophisticated and expensive." On the other hand, Beggs adds, "there is a lot of research that could be done on the cheap.... We've got to get better at this; scientists need to be able to get into space on a short time fuse and relatively inexpensively."

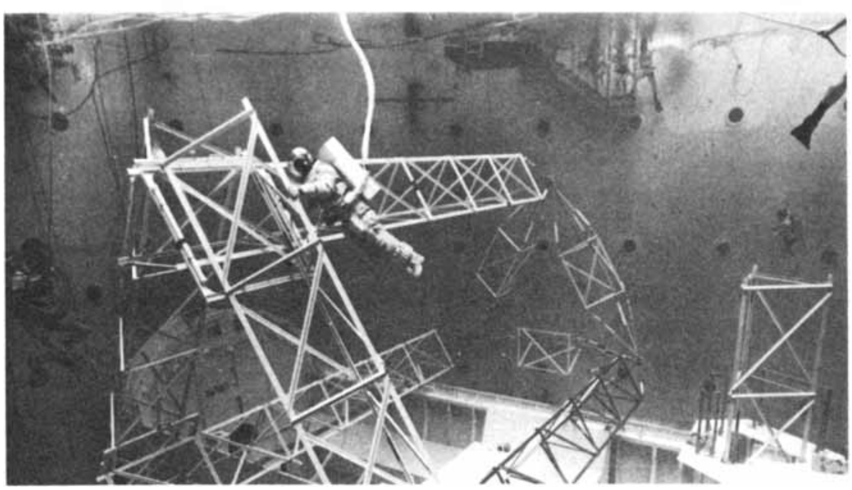
The tendency to emphasize large projects is not the only point on which NASA has been criticized. A report released in November, 1984, by the congressional Office of Technology Assessment (OTA) urges strongly that NASA and the Federal Government thoroughly reconsider the entire structure of the current space program.

The OTA report was originally commissioned as a study of NASA's proposed space station. According to the OTA, it is not yet possible to judge whether the space station NASA has proposed will help the nation to meet its goals in space because no such goals have been defined. The report states that space science and commercialization will soon enter a period of rapid development; therefore the nation must now consider what its objectives are and what role NASA should play in meeting these objectives.

Beggs disagrees with the OTA on the subject of goals. He asserts that NASA does indeed have long-range goals. The major goals, says Beggs, must be those that are politically acceptable. Within such goals there is room to develop others: "the highly visible goals provide a focus" around which NASA can develop its capabilities. Beggs believes NASA's goals for the space station have been articulated very clearly: it will provide a capability for routine operations in space.

The OTA report does not question the need for a capability; it stresses that the nation must determine what it wants to do before any facilities to provide such a capability are designed.

According to the report, time has become a critical factor. Since the space-shuttle development program will be coming to an end in the next two years, NASA must begin other ambitious projects within that time or risk losing many of its best engineers and managers to private industry or to defense projects. The U.S., the report states, has only two years within which to define its goals in space and initiate projects to meet them.



*Rehearsing for NASA's future: underwater simulation of deep-space construction*

## *Simplified Simplex*

For almost 30 years oil refineries, paper manufacturers, food distributors, steelmakers and metal fabricators have depended on a mathematical method called the simplex algorithm for determining the optimum way of deploying their resources. The algorithm has also been applied to the distribution of petroleum products, the scheduling of flight crews, the routing of long-distance telephone calls and the scheduling of time on communications satellites. Its introduction has routinely saved millions of dollars on operating costs within a single company. Now Narendra Karmarkar of AT&T Bell Laboratories has invented a new algorithm that appears to solve such problems much faster than the simplex does.

Karmarkar's algorithm, like the simplex, can be applied to any circumstance that can be considered a problem in linear programming. The advantage of such algorithms is that they identify the optimum solution to any linear-programming problem without having to explore all the feasible ways of deploying resources, which is usually impossible.

Karmarkar's algorithm appears to be the first significant practical advance in solving such problems since the simplex algorithm was invented in 1947. Five years ago four mathematicians in the U.S.S.R. devised an algorithm called the ellipsoid method that is faster than the simplex method for certain rather artificial problems. In practice, however, such problems almost never arise, and almost all practical problems are solved many times faster by the simplex method than they are by the ellipsoid.

Karmarkar's algorithm seems to win on both counts. It appears to improve on the practical performance of the simplex algorithm, and its theoretical, "worst case" performance is better than that of the ellipsoid method.

The general nature of problems in linear programming can be illustrated by the predicament of an oil refiner who can make only two grades of fuel: high-octane and low-octane. Suppose the refiner combines both low- and high-grade crude oil to make the low-octane fuel, whereas only high-grade crude is needed for the high-octane fuel. If the current market prices make high-octane fuel more profitable than low-octane fuel, the refiner might choose to make as much high-octane fuel as possible.

Oddly, such a strategy might not be the most profitable one. The costs of storing the current supply of low-grade crude left over from the previous month's heavy demand for low-octane fuel could wipe out the profits. It could well be more profitable to make more effective use of the low-grade crude on hand by allocating part of the current production to low-octane fuel.

All these constraints can be represented as a system of linear inequalities. The inequalities can then be plotted on a graph whose horizontal axis represents, say, the output of low-octane fuel and whose vertical axis represents the output of high-octane fuel. Each point on the graph represents a production program that gives some mix of low- and high-octane fuel. The set of all points that satisfy the constraints on the raw materials forms a polygon in the plane.

There is some profit to the refiner associated with every point of the polygon. It turns out that the maximum profit to the refiner is associated with one of the vertexes of the polygon. It is this point that is sought by any linear-programming algorithm.

The simplex algorithm begins its search at an arbitrary vertex and seeks to reach the optimum vertex by moving from vertex to vertex along the edge of the polygon. Karmarkar's new method takes a more direct approach. It picks a starting point on the polygon and then attempts to improve the profit by taking a short cut through the polygon rather than merely moving along some edge. The outcome of the search is the optimum vertex, as it is with the simplex algorithm, but the method is so much faster than the simplex that it can be applied to much larger and more complex problems. The Karmarkar algorithm is now being translated into high-speed machine language so that its full capabilities can be more fairly compared with those of the simplex.

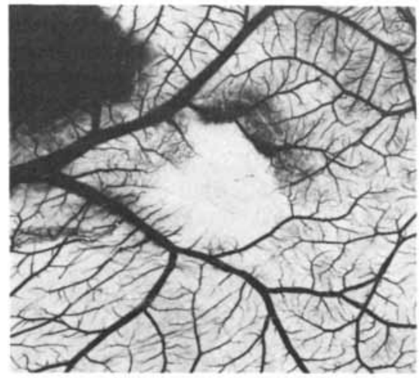
## Blood Donor

Back in 1961 it occurred to Judah Folkman, who was then working at the Naval Medical Research Institute, that a tumor might have to develop its own network of blood vessels in order to grow beyond a millimeter or two in diameter. If that were so, inhibition of angiogenesis (blood-vessel formation) might inhibit tumor growth.

For the past 20 years, as an investigator at the Harvard Medical School, Folkman has sought conclusive evidence that solid-tissue tumors promote the growth of their own vascular system by secreting a tumor angiogenesis factor, which stimulates the proliferation of the endothelial cells that form capillaries. Within the past year he and his colleagues have purified such a factor, and in doing so they may have discovered a way to inhibit its activity.

To test his hypothesis Folkman and his associates developed new techniques for growing tumors in an environment isolated from blood vessels (inside a filter chamber or in avascular tissues) and for growing endothelial cells in culture. It took several years to establish clearly that tumors do secrete a diffusible endothelial-cell growth factor that stimulates angiogenesis. For a long time, however, it was not possible to isolate a pure factor.

The path to purification was opened when workers in Folkman's laboratory found that heparin promotes migration of endothelial cells and hence increases the rate of angiogenesis. Heparin, a polysaccharide (complex sugar) present in many tissues, is exploited as a drug for inhibiting coagulation. Yuen Shing and Michael Klagsbrun thought heparin might promote an-



*The inhibitory effect of heparin-cortisone*

giogenesis by increasing the concentration of a growth factor at the surface of endothelial cells. To take advantage of such a factor's putative affinity for heparin they coated the grains of gel in a filtration column with heparin and passed through the column a partially purified factor isolated from a chondrosarcoma, a cartilage-cell tumor.

Most of the crude factor passed through the column. What stuck to the heparin-coated gel was separated from the heparin and subjected to gel electrophoresis, which sorts proteins according to size. It was found to be a single small protein, and a potent promoter of both endothelial-cell growth and angiogenesis.

Shing and Klagsbrun think the chondrosarcoma-derived substance may be a growth factor normally present in cartilage, where it may stimulate the vascularization attending the conversion of cartilage into bone. In other words, a tumor angiogenesis factor is probably not peculiar to tumor cells; it may be a normal growth factor that is turned on at the wrong time or is kept on too long.

Heparin turns out to have a role in inhibiting angiogenesis as well as promoting it. Testing the enhancement of angiogenesis on a chick-embryo membrane, Stephanie Taylor added the steroid hormone cortisone to heparin in an effort to prevent inflammation. Angiogenesis ceased; embryonic blood vessels regressed from a disk impregnated with the two drugs. Neither drug alone had any such effect. When it is given orally to mice, the heparin-cortisone combination causes some kinds of tumor to shrink.

## Lost Twin

The unification of the weak nuclear force and the electromagnetic force, two of the four fundamental forces in the universe, leads naturally to a more ambitious question: How are gravity and the strong nuclear force to be included in a unified scheme? One prominent answer in recent years has predicted the existence of a twin universe of new particles. Each hypothetical particle is the counterpart of one of the particles whose existence is now recognized by experimental physics. The theory is called supersymmetry, and its experimental consequences are being systematically explored. The latest evidence is not encouraging.

Workers at the MARK-J detector of the PETRA accelerator at the Deutsches Elektronen-Synchrotron (DESY) looked for evidence of supersymmetric twins of the $W^+$, $W^-$ and $Z^0$ vector bosons. The three particles were discovered at CERN, the European laboratory for particle physics, in 1983, and their measured masses make it possible to give a more reliable estimate of the masses of their hypothetical supersymmetric twins. Making certain assumptions about the decay products of the twins, the investigators could find no evidence for the supersymmetric particles.

More precisely, the results imply that such particles, if they exist at all, must be quite massive, although not as massive as their recently discovered counterparts. The twin of the $Z^0$ must be heavier than 35 GeV (billion electron volts), and the twins of the two $W$ particles must each weigh at least 25 GeV.

## *X-radiant*

An X-ray laser could enhance the precision of X-ray spectroscopy, produce holograms of cellular and molecular structures in organisms and etch ultrafine circuit patterns on semiconductor chips. Converging developments are bringing such a device closer to reality.

At the Lawrence Livermore National Laboratory a team headed by Dennis L. Matthews has created plasmas that generate laser beams of soft (long-wavelength) X rays; a group at the Princeton Plasma Physics Laboratory led by Szymon Suckewer has reported similar results. Meanwhile Troy W. Barbee, Jr., of Stanford University has developed mirrors that can reflect X rays.

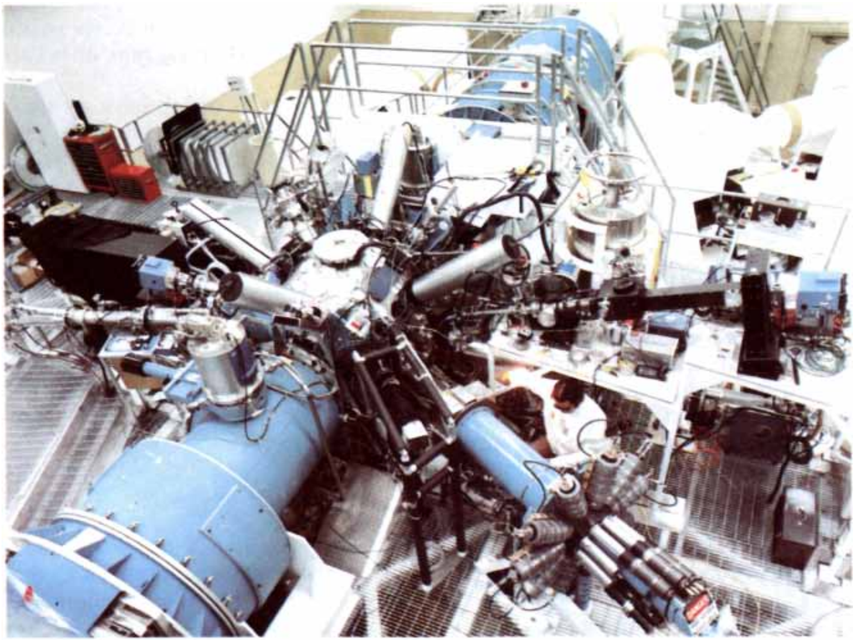
The difficulty of regulating energy emission at X-ray wavelengths has hindered efforts to develop such a laser. To raise an atom or an ion to the energy level necessary for it to release an X-ray photon it must be "pumped" with enormous amounts of energy. Such atoms are elevated to quantum states so high that they rarely stay there long enough to allow lasing, a cascade of stimulated emission in which photons trigger the release of other photons of the same wavelength and phase. Instead the electrons tend to relax spontaneously into lower energy levels, randomly emitting X rays.

The Livermore workers identified yttrium and selenium as substances in which sufficiently high quantum states can be maintained. The laboratory's giant Novette laser, which can produce visible light at an intensity of $5 \times 10^{13}$ watts per square centimeter, provided the energy needed to reduce tenuous films of the metals to a plasma. Its pulses touched off directed bursts of X rays lasting 250 trillionths of a second, bearing all the hallmarks of laser radiation.

The X rays from the selenium plasma were about 700 times more intense than spontaneous emissions and fell at specific wavelengths: 206 and 209 angstrom units. The yttrium plasma lased at a wavelength of 155 angstroms. The intensity of the X-ray bursts increased exponentially in relation to the length of the plasma column, proof positive of lasing.

Using a carbon plasma, generated with infrared pulses from a carbon dioxide laser, the Princeton team has achieved X-ray lasing at a wavelength of 182 angstroms. The group has so far not been able to determine whether the amplification increases exponentially with the length of the plasma column.

Mirrors that reflect X rays promise considerable gains in intensity. In optical lasers mirrors at each end reflect the pulse back and forth through the medium, greatly intensifying the beam. Until recently X-ray mirrors were unknown. Now Barbee has devised surfaces consisting of layers of molybdenum, 35 angstroms thick, separated by 60-angstrom layers of silicon. Each layer of molybdenum reflects only a small number of the incident X rays. The rest pass through to the next layer; because of the precise spacing of the layers, however, the feeble reflections from each one interfere constructively, amplifying the reflected beam so that an efficiency approaching 70 percent is achieved.



*The Novette laser's instrument-flanked target chamber, where X-ray lasing was induced*

## *Shredded Knapsack*

Two years ago Ralph C. Merkle of ELXSI in San Jose, Calif., offered $1,000 to anyone who could break the most general version of the knapsack code, a coding system he had helped to devise. Merkle was said to be upset by the widespread publicity given to Adi Shamir of the Weizmann Institute of Science in Israel when Shamir cracked only one version of the code (see "Science and the Citizen," SCIENTIFIC AMERICAN; August, 1982). Now Merkle has paid up. Ernest F. Brickell of the Sandia National Laboratories has broken the code and has thereby shown that the general version of Merkle's system is insecure.

The knapsack code, which has never been used commercially, is based on a venerable combinatorial puzzle called the knapsack problem. Merkle and Martin E. Hellman of Stanford University expanded the problem into a cryptographic system in 1977.

The problem is simple to state, but it can be exceedingly difficult to solve.

Suppose a knapsack is to be filled to capacity with objects selected from a large stock. The weight of each object is known. The problem is to determine which set of objects, if any, will exactly fill the knapsack. The capacity of the knapsack is typically given by a number 62 digits long, and the weight of each object is given by a 60-digit number. The sum of roughly 100 such numbers selected from the stock is equal to the 62-digit number.

If weights of such magnitude were selected arbitrarily, the code would be impossible to break even by the intended receiver of the message. Instead the weights are selected from a stock with special properties, which considerably simplifies the problem. The security of the coded messages then depends on disguising the weights in such a way that to the eavesdropper they are indistinguishable from a set of arbitrary weights.

Each weight is multiplied by a large random number, the product is divided by a second large random number and the remainder after the division is the weight that is placed in the general stock. The procedure can be repeated, disguising the disguise.

In 1983 Leonard M. Adleman of the University of Southern California and Jeffrey C. Lagarias of AT&T Bell Laboratories discovered attacks on the doubly iterated system. Lagarias, his colleague Andrew Odlyzko and Brickell then extended the attack to an entire class of knapsacks.

## *Special Delivery*

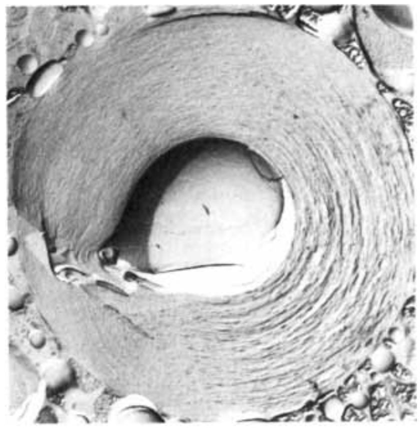
Efforts to administer synthetically produced peptide drugs for the treatment of metabolic diseases face a daunting obstacle: the body's own enzymes quickly cut such molecules into their component amino acids. Hence the therapeutic effect of a single injection is limited to a comparatively short time span, often less than an hour. A drug-delivery system that could gradually release the peptide molecules within the body would solve the problem. Recent developments have revived interest in a long-standing candidate for the job: the miniature spheres of artificial cell membrane known as liposomes.

The first liposomes were made more than two decades ago in the laboratory of A. D. Bangham of the British Agricultural Research Council's Institute of Animal Physiology in Cambridge. Bangham and his co-workers observed that when certain lipids (the basic building blocks of biological membranes) are dispersed in water, they tend to form closed vesicles. Most of these structures consist of many two-ply layers of lipid molecules shaped into concentric spherical shells. Exposing the solution to high-frequency sound waves, it was later found, further reduces the size of the lipid vesicles (dubbed liposomes) to more uniform microscopic dimensions.

It was recognized fairly early that drugs and other biologically active compounds could be trapped in either the lipid bilayers or the aqueous spaces separating them. The exploitation of this promising carrier technique was hampered, however, by the fact that there seemed to be no economic way to produce liposomes that were both stable and capable of efficient entrapment. As a result liposomes were long regarded as a laboratory curiosity, useful chiefly as models for the study of biological membranes.

Nevertheless, work on various aspects of liposome technology went forward. In 1981 a new commercial venture, the Liposome Company, Inc., was formed to exploit the growing body of knowledge about liposomes. The company initially concentrated on the three main problems of the carrier system: entrapment, stability and economy. Two years ago investigators at the company's laboratory in Princeton reported that they had succeeded in developing a process to produce stable, low-cost liposomes with a high entrapment efficiency.

Since then it has been shown that the new liposome, made largely of a type of lipid (lecithin) derived from egg yolks, is capable, when it is injected into muscle, of releasing its drug contents over a long period (up to 20 days) as the lipid bilayers are slowly metabolized. Late last year the Liposome Company announced that it had entered into a long-term, multimillion-dollar agreement with the government of the Walloon region of Belgium, a Belgian venture-capital firm and the Catholic University of Louvain. The aim of the joint project is to combine an advanced drug-targeting technique developed at the Belgian university with the liposome drug-delivery system to treat malaria and other diseases that affect the liver, such as hepatoma and hepatitis.



*A freeze-fractured liposome*

## *MacIsland*

When Macdonald Seamount was surveyed in 1975, its peak was about 49 meters below the surface of the Pacific. Since then this undersea volcano about 1,000 miles southeast of Tahiti has been erupting noisily and apparently growing fast. Writing in *Geophysical Research Letters,* Jacques Talandier of the Laboratoire de Géophysique in Tahiti and Emile A. Okal of Northwestern University speculate that Macdonald may emerge as an island within the next few decades.

Since 1977 seismic stations on several Polynesian islands have detected a total of 12 seismic "swarms" that Talandier and Okal say can be clearly attributed to eruptions at Macdonald; the volcanic activity reached its greatest intensity in 1983. In July of that year divers from a French naval escort vessel explored the plateau around the summit and found tangible evidence of recent volcanism: lava spatter cones, free of algae, flanking a fissure with fresh walls. Soundings indicated the summit of the volcano had risen to just 27 meters below sea level.

Macdonald lies at the southeastern end of the Austral Islands, a chain that conspicuously parallels the Hawaiian Islands as well as the closer Tuamotu group. According to plate-tectonic theory, all three chains were formed by the northwestward drift of the Pacific plate over stationary hot spots: plumes of molten rock welling up from deep in the mantle. Like Hawaii and Pitcairn Island (the youngest member of the Tuamotu group), Macdonald Seamount is the active expression of a hot spot—and of the explanatory power of plate tectonics.

## *Dark-Age Diet*

What did the people of Europe eat during the Middle Ages? It has been thought that the medieval diet was poor in protein and heavy in starch, much of it derived from bread. The excavation of the town of Dorestad in the Netherlands, however, is providing a rich haul of artifacts that contradict the prevailing theory.

In A.D. 800 Dorestad, with a population of 10,000, was the largest town west of Constantinople. Situated at the confluence of the Rhine and Lek rivers, the settlement was an important trading center: it supplied the Rhine-land with imports from Britain and Scandinavia. Piers lined the rivers, and great warehouses stood behind the piers. A large village occupied the center of the site.

Analysis of the organic remains from Dorestad by Wietske Prummel of the Dutch State Institute of Archaeology shows that the inhabitants' diet was rich in protein. Furthermore, Prummel has estimated that about 30 percent of the protein came from animal sources. The average household had 10 cows, 10 sheep, 10 pigs, 20 chickens and a goose. Since Dorestad was a trading settlement, the diet may have been enriched further by imports.

Bones from Dorestad show that the individual animals were large and the herds well managed. In the journal *Nature* Richard Hodges of the University of Sheffield writes that along with findings from Southampton in England and other villages throughout Europe "animal bones from Dorestad... challenge the primeval picture of early medieval life."

## *E Is for Enceladus*

So many of the outer planets have been found to have rings that these dramatic features have become as much an astronomical commonplace as moons. Rings are generally thought to result from collisions between satellites, fragmentation of a single body through the shearing force of the planet's gravitational attraction, or the accretion of material left over as the planet formed. Not so Saturn's *E* ring. This faint feature of the planet may have a most unusual origin and composition.

Detailed analysis of data sent back by *Pioneer 11*, *Voyager 1* and *Voyager 2* has made it possible to propose that the ring consists of minute particles spewed out by volcanoes on Enceladus, one of Saturn's 17 known moons, which resides in the *E* ring.

Writing in *Journal of Geophysical Research*, Kevin D. Pang, Charles C. Voge, Jack W. Rhoads and Joseph M. Ajello of the Jet Propulsion Laboratory note the ring is brightest at the orbit of Enceladus and dims exponentially with increasing distance from the orbit. That is enough to have suggested to many investigators that Enceladus, a moonlet only some 500 kilometers in diameter, supplies the ring's component particles. Analysis of the optical properties of the ring indicates it is composed of ice spheres from two to 2.5 micrometers (thousandths of a millimeter) in diameter.

The authors argue that the particles come from volcanic eruptions on Enceladus, where tectonic activity has been reported. Because its orbit is slightly elliptical, the moonlet oscillates in Saturn's gravitational field; the resulting tidal forces could keep water in the interior melted and provide the energy for episodic eruptions.

## *Enzymes' Mettle*

Atoms of zinc are essential components of many protein-digesting enzymes: they may have both structural and catalytic functions. John M. Prescott of Texas A&M University, Frederick W. Wagner of the University of Nebraska and Barton Holmquist and Bert L. Vallee of the Harvard Medical School have found that substituting other metals for zinc in one such enzyme dramatically enhances its activity.

The workers studied an enzyme abundantly secreted by a marine bacterium, *Aeromonas proteolytica*. Joseph R. Merkel of Lehigh University isolated the bacterium while searching for an organism that would digest cellulose. Its enzyme had no effect on cellulose; instead it readily hydrolyzed proteins.

As with other such enzymes, the action of the substance seemed to depend on constituent atoms of zinc. When the workers removed all but 5 or 6 percent of the zinc from the enzyme, its efficacy dropped correspondingly, to 5 or 6 percent of normal. But when the researchers replaced the zinc with proportionate amounts of copper or nickel, the enzyme's activity was multiplied from six to 25 times.

The workers cannot account for the effect. Prescott guesses that the alien metal changes the conformation of the enzyme molecule, thus altering the way it interacts with the substrate. A clearer answer may emerge from X-ray crystallographic studies now under way at the University of Alberta.

Given the ubiquity of zinc-containing enzymes in biological systems, including those of the human body, a way to enhance their function could have interesting applications.

## *All in the Timing*

Almost all multicellular organisms, no matter how complex, originate in a single cell. The mature organism may include billions of specialized cells. Hence it is clear that for maturation to occur smoothly the timing of cellular differentiation must be precisely controlled.

Until recently the nature of the genetic mechanism that controls timing was little understood. The discovery of genes that influence the pace of development of specific cell lines is beginning to shed light on this question. In addition the discovery of such genes could provide a molecular basis for the evolution of new species.

Genes that have a role in determining the relative timing of developmental events have been named heterochronic genes by their discoverers, Victor Ambros of Harvard University and H. Robert Horvitz of the Massachusetts Institute of Technology. Ambros and Horvitz coined the term to account for the effects of mutations they induced experimentally in a small transparent roundworm called *Caenorhabditis elegans*.

The effects of heterochronic mutations can vary widely. Mutations that are dominant with respect to the wild-type, or normal, gene cause some cell lines to repeat the patterns shown by their predecessors instead of developing into more highly differentiated descendant cells. Recessive mutations, in contrast, cause cells to display prematurely the developmental patterns normally shown by their more advanced descendants. By observing the effect of such mutations the investigators can attempt to infer the function of the wild-type gene.

For example, the normal development of *C. elegans* includes several larval stages and one adult stage. The cuticle, or external covering, of the larval stages is quite distinct from that of the adult stage. In animals with one type of dominant heterochronic mutation the cuticle-forming cells in the adult replay the genetic program of the larval stages. As a result sexually mature adults have a larval cuticle. In animals with a recessive mutation the adult cuticle appears prematurely at a larval stage.

The emergence of new species that show a combination of mature and immature anatomical structures could be due to mutations in heterochronic genes. The Mexican axolotl, for instance, resembles the larva of the salamander but is sexually mature. In an article in *Science* describing their findings Ambros and Horvitz note that the appearance of the axolotl as a species could have resulted from a heterochronic mutation that causes the sexual tissues to mature while the development of the rest of the body is retarded.

The temporal transformations of cell fates induced by heterochronic mutations are analogous to the spatial aberrations induced by homeotic mutations, which can cause a *Drosophila* embryo to grow an extra leg on its head. The mechanisms by which homeotic and heterochronic genes operate in the normal organism are currently the subject of intensive investigation in several laboratories.

# 8087 AND 80287 TECHNICAL TOOLS

**87FFT™** performs Forward and Inverse FFTs on real and complex arrays which occupy up to 512K bytes of RAM. Also does convolutions, auto correlations, hamming, complex vectormultiplication, and complex to radial conversions. Callable from MS Fortran or 87BASIC/INLINE ..... **$150**

**87FFT-2™** performs two-dimensional FFTs. Ideal for image processing. Requires 87FFT...**$75**

**MATRIXPAK™** manages a **MEGABYTE!** Written in assembly language, our runtime package accurately manipulates large matrices at very fast speeds. Includes matrix inversion and the solution of simultaneous linear equations. Callable from MS Fortran 3.2, 87MACRO, 87BASIC/INLINE, and RTOS ........ each **$150**

**DATA ACQUISITION PACKAGE**
Interactive, user-oriented language which allows the acquisition and analysis of large data streams .................................. **CALL**

**GRAPHICS PACKAGES**
Energraphics (stand alone) ................. 295
Grafmatic for MS Fortran or Pascal ........ 125
Plotmatic for Grafmatic .................... 125
Halo for Basic, C or Fortran .......... each 150

**OTHER TOOLS**
Alpha Software ESP ........................ 595
Borland Sidekick, Toolbox, or Graphics ...... 45
COSMOS Revelation ....................... 850
Lattice C .................................. 299
PSI MATHPAK ................................ 75
smARTWORK ............................... 895
SPSS/PC ................................... 695
STSC APL★PLUS/PC ......................... 475

**RTOS - REAL TIME OPERATING SYSTEM**
RTOS is a multi-user, multi-tasking real time operating system. It includes a configured version of Intel's iRMX-86, LINK-86, LOC-86, LIB-86, OH-86, and MicroWay's 87DEBUG. Runs on the IBM-PC, XT, PC-AT and COMPAQ ........................ **400**

**INTEL COMPILERS**[1]
FORTRAN-86 ................................ 750
PASCAL-86 ................................. 750
PL/M-86 ................................... 500
87C (LATTICE/MICROWAY) .................... 750
ASM-86 .................................... 200

**URS™ - Universal Run Time System**[1]
Generates programs with the Intel compilers which run on other operating systems. MS-DOS version is included with RTOS.
Xenix-286 Version .......................... 300

**SoftScope Symbolic Debugger**[1] **500**

[1]Requires RTOS or iRMX-86. All Intel compiler names and iRMX-86 TM Intel Corp.

**87BASIC/INLINE™** converts the output of the IBM Basic Compiler into optimized 8087 inline code which executes up to seven times faster than 87BASIC. Supports separately compiled inline subroutines which are located in their own segments and can contain up to 64K bytes of code. This allows programs greater than 128K! Requires the IBM Basic Compiler and Macro Assembler. Includes 87BASIC ............ **$200**

**PC AT and 86-310 DRIVES**
30 MEGABYTE WINCHESTER ........... **2000**
53 MEGABYTE WINCHESTER ........... **2600**
SYQUEST FIVE MEGABYTE ............. **950**

**HARDWARE AND LANGUAGES**

**8087-3 5mhz ........ $149**
Including DIAGNOSTICS and 180-day warranty
For IBM PC and compatibles

**8087-2 8mhz ....... $275**
For Wang, AT&T, DeskPro, NEC, Leading Edge

**80287-3 5mhz ..... $275**
For the IBM PC AT

**64K RAM Set ............ $30**

**256K RAM Set ........... $195**

**128K RAM Set PC AT .$225**

**NUMBER SMASHER™...CALL**
10mhz 8087 coprocessor board for the IBM PC

**FORTRAN and UTILITIES**
Microsoft Fortran 3.2 ...................... 239
IBM Professional Fortran ................... 595
Intel Fortran-86[1] ........................ 750
FORLIB+ ..................................... 65
STRINGS and THINGS .......................... 65

**BASIC and UTILITIES**
IBM Basic Compiler ........................ 270
87BASIC/INLINE ............................ 200
Summit BetterBASIC™ ....................... 175
Summit 8087 Module .......................... 87

**MACRO ASSEMBLERS**
IBM Assembler with Librarian ............... 155
87MACRO ................................... 150

**PASCAL**
Microsoft Pascal 3.2 ....................... 209
Borland Turbo ................................ 45
Turbo with 8087 Support ..................... 85

**Micro Way** P.O. Box 79 Kingston, Mass. 02364 USA (617) 746-7341

**You Can Talk To Us!**

**NO CHARGE FOR CREDIT CARDS**
**ALL ITEMS IN STOCK**
**CALL FOR COMPLETE CATALOG**

---

# SCIENTIFIC AMERICAN

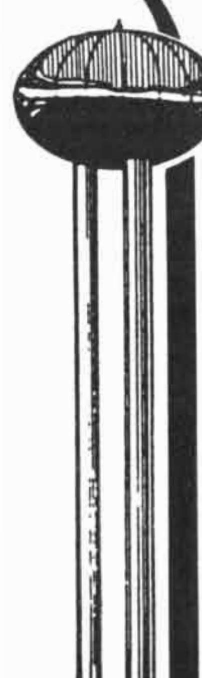
**is now available to the blind and physically handicapped on cassette tapes.**

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

---

# Explore Tomorrow's Automotive Frontiers Now With General Motors



General Motors' Advanced Product and Manufacturing Engineering Staff has immediate challenging opportunities in Factory of the Future Engineering.

**FACTORY OF THE FUTURE ENGINEERING**
Coordinate the development of the information and control systems for the Factory of the Future implementations. Investigate the potential role of artificial intelligence in these control systems. Assist Factory of the Future sites in implementing these information and control systems and in evaluating the technical innovations of vendors. Candidates should possess BS/MS in Engineering or Mathematics.

Candidates must have U.S. Citizenship or Visa status which permits them legally to accept permanent employment under U.S. Immigration laws.

Exploring new automotive frontiers with GM means virtually unlimited opportunity to realize your full career potential, plus the outstanding salaries and benefits only the leader in its industry can offer. To receive prompt, confidential consideration, please submit your resume with salary history/requirements to:

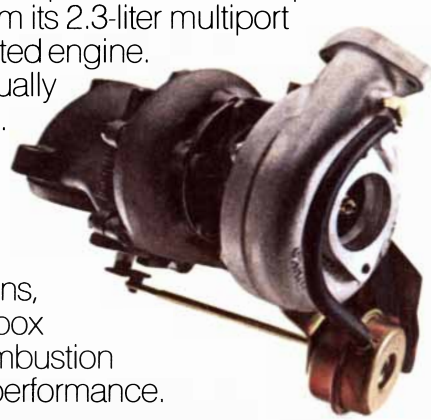
GM

**Technical Staffs Personnel - 75P**
**GENERAL MOTORS TECHNICAL CENTER**
**Engineering Building North**
**30200 Mound Road**
**Warren, MI 48090-9010**

An Equal Opportunity Employer

# Are we the on a performance ca this way?

XR-7's turbo pulls a 60% horsepower boost from its 2.3-liter multiport fuel-injected engine. With virtually no lag.

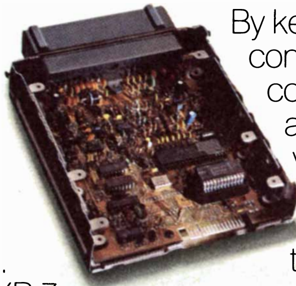


Not everyone shares our view of what makes a performance car a performance car.

Some offer merely the momentary elation of acceleration.

Others offer little more than the status of owning them.

But the Mercury Cougar XR-7 was built with the belief that the truest test of a performance car lies not in how well it does one thing. But in how many things it does well.



By keeping constant command over an orchestra of vital engine functions, this curious little box can maximize combustion to give you better performance.

# y ones who think
# r should be made

This is something you won't find on many cars.

Two extra shocks.

For more control in both turning and acceleration.

Plus sticky Goodyear Eagle tires, a turbocharged engine, and sport suspension. Because we don't believe in giving you the power to perform without also giving you the tools to use it wisely.

No special effects. No tricks. Just the facts.

To us, an instrument panel is an opportunity to say as much as possible in the least possible time.

When we designed articulated seats to adjust precisely to give you thigh, hip, and lumbar support, it wasn't just because we thought they'd look good.

If you believe the ability of a performance car to perform in more ways than simply straight-ahead acceleration is what makes it worth the sticker price, we suggest you call us at 1-800-MERCFAX for a free 1985 Cougar XR-7 catalog. We may have made a performance car the way you were hoping someone would.

# Mercury Cougar XR-7 Turbo

# Alzheimer's Disease

*It brings dementia and slow death to more than 100,000 Americans a year. No one knows its cause or how to stay its inexorable course. Investigators are focusing on six conceptual models of the disease*

by Richard J. Wurtman

Every year several hundred thousand Americans begin to lose their ability to remember whether they have turned off the stove or locked the front door. It becomes hard for them to name familiar objects, find the right word or balance a checkbook. They continue to look well, and they exhibit neither overt neurological deficits nor signs of serious depression, stroke, brain tumor or any other specific disease likely to give rise to such symptoms. Yet these people are very sick. Within three to 10 years they will be severely demented: in effect, deprived of reason. They will be unable to speak or think or take care of themselves, and in time they will die of some complication that afflicts bedridden patients.

Most of these people are elderly, but their symptoms are not the result of "aging" as such. They have a specific sickness: Alzheimer's disease. The diagnosis can only be inferred during the patient's lifetime; no unique pattern of behavioral abnormalities has been established and there is no satisfactory laboratory test short of a brain biopsy. An autopsy, however, shows highly characteristic pathologic changes in the brain.

There is a loss of neurons (nerve cells), particularly in regions essential for memory and cognition, or thought processes. There are accumulations of twisted filaments (neurofibrillary tangles) and other abnormal structures within neurons. There are amorphous aggregates of protein (amyloid) adjacent to and within blood vessels. And there are scattered focuses of cellular debris and amyloid called neuritic plaques. (Such changes are also seen in the brain of some old people who are not demented. In these people, however, usually not all the changes can be observed in a single case, nor are they present in such abundance.) There is also a significant loss of neurons in certain more primitive regions at the base of the brain, with consequent reduction in the amount of the neurotransmitters (chemical messengers), notably acetylcholine, normally released from the terminals of those neurons in higher brain centers.

Why is it that such a devastating condition attracted little attention from physicians until recently and that its prevalence was not recognized? To the medical scientist a disease is a clinical and pathological entity: a characteristic and usually progressive set of changes in the appearance and functioning of the patient's body and in the gross or microscopic structure of affected organs and tissues. As long ago as 1907 the German neurologist Alois Alzheimer recognized this disease as a specific clinical and pathological entity that seemed to afflict a few people in their forties and fifties.

For a long time the disease named for Alzheimer was considered to be a specifically presenile dementia; elderly people with comparable symptoms were said to suffer from "senility" or "hardening of the arteries." Within recent years it has become clear that the brain of most old people with dementia shows all the characteristic signs of Alzheimer's disease if it is examined at autopsy; indeed, the abundance of such signs is correlated with the degree of mental deterioration. Now it is estimated that Alzheimer's disease afflicts from 1.5 to two million Americans and that at least 100,000 of them die of it every year. Yet no one knows what causes the disease, how its characteristic changes are brought about or how to treat it.

How can a disease be transformed from one that can only be described to one that is understood and can be treated? One way to approach this goal is to propose conceptual models based on the symptoms and pathological findings and on the processes that might give rise to them. Each model is then tested and either discarded or refined until its validity is established, usually by its ability to predict effective treatment. Six models now underlie most research on Alzheimer's disease. They arise from an equal number of hypotheses accounting for the loss of neurons.

One hypothesis starts with the observation that a particular neurotransmitter, acetylcholine, appears to be in short supply in the brain of Alzheimer's disease patients; it proposes explanations for the selective loss of the neurons from whose terminals acetylcholine is released. The other hypotheses are more general. They propose respectively that the death of neurons is caused by faulty genes, by abnormal accumulations of proteins, by an infectious agent, by an environmental toxin or by inadequate blood flow and energy metabolism. Each model is supported by some observational or experimental evidence—and each seems to be contradicted by other evidence. I shall consider each of these models, leaving for the last the acetylcholine model, which has attracted the most interest and has given rise to most attempts at therapy.

## The Genetic Model

There are families in which the incidence of Alzheimer's disease is unusually high; that fact alone constitutes strong evidence for the involvement of a heritable factor in at least some forms of the disease. Is it a faulty gene (or more than one)? Does some factor in the genetic makeup render a person particularly vulnerable to some environmental factor?

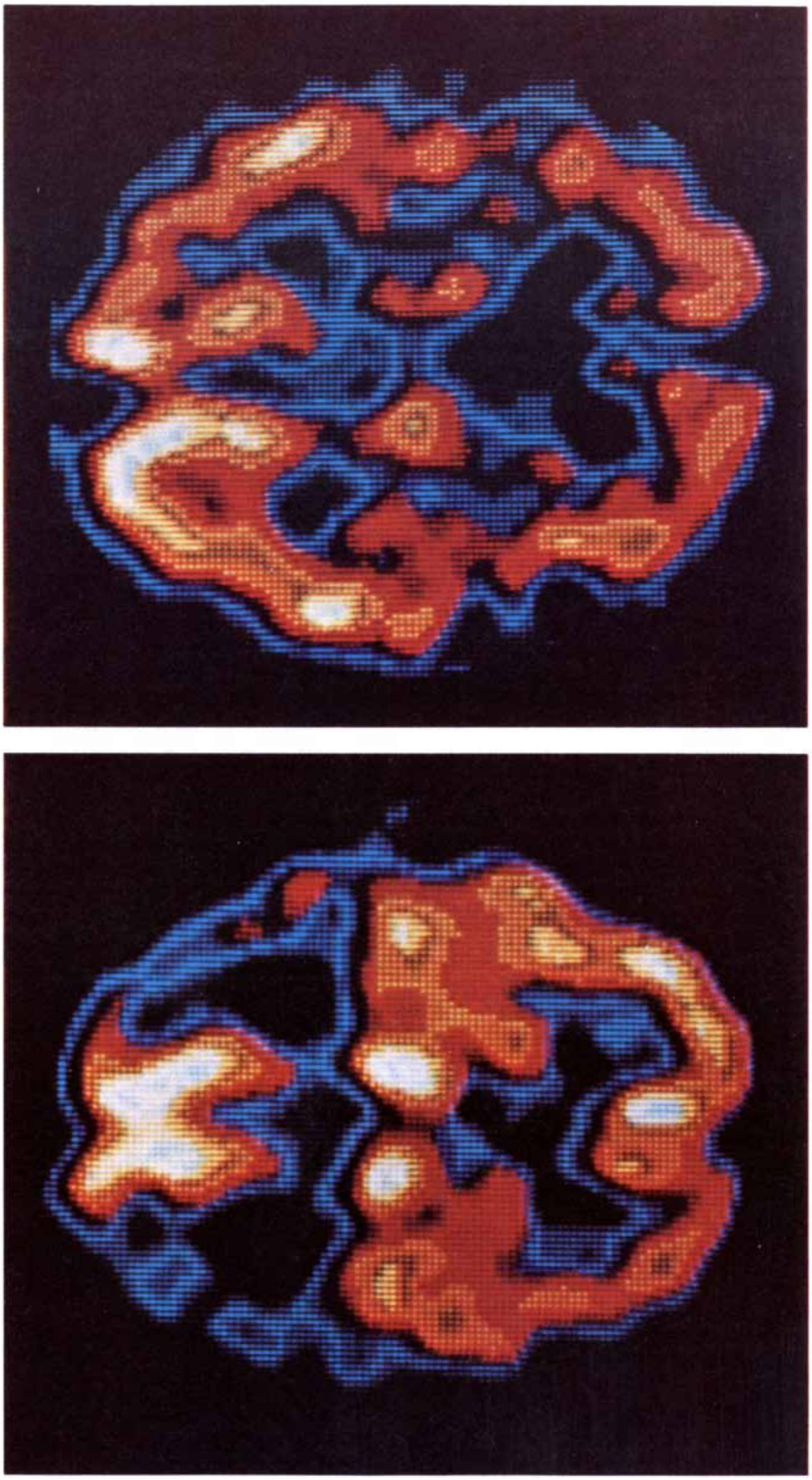
A typical genetic disease stems from an inborn error of metabolism: an abnormality in the genetic material, DNA, impairs the ability of cells to make a particular protein. If the protein is an enzyme, it can often be identified by a finding that the substance whose conversion the enzyme ordinarily catalyzes is present in excess (because the enzyme-catalyzed reaction has not taken place) and that the usual product of the reaction is in correspondingly short supply. A genetic disease can also manifest itself as an abnormal chromosome; that is the case in Down's syndrome, or mongolism.

If there is a deficiency of an important enzyme, metabolic and clinical abnormalities are commonly present from birth. Still, certain diseases are known to have a genetic origin and yet to be "silent" during the first decades of life. Huntington's disease, the manifestations of which include abnormal movements, progressive dementia and the loss of particular groups of brain neurons, usually becomes clinically apparent only after the age of 40. How might a genetic error, which is present by definition from the moment of conception, not reveal iself until after many years?

Perhaps aging further impairs some genetically compromised physiological system such as the immune system; perhaps it worsens an existing deficiency in the manufacture of some essential protein (such as a growth factor sustaining cholinergic neurons). In either case one should be able to identify a genetically deficient protein in the brain of some members of families with a high incidence of the disease, and the deficiency should be found to worsen with age.

Families have been identified in which 10 or more members, representing four or five generations, have developed a dementia of the Alzheimer's type. In such cases it appears that an aberrant gene is transmitted as an autosomal dominant: a gene that affects both sexes equally and that need be inherited from only one parent to have full effect. A genetic component can be indicated by large-scale studies of family trees even in families in which the disease appears only sporadically. In one such study Leonard L. Heston of the University of Minnesota Medical School either examined or obtained information about the parents, siblings and second-degree relatives of 125 patients, in different families, who died in Minnesota hospitals or nursing homes from 1952 to 1972 and showed adequate evidence at autopsy for a diagnosis of Alzheimer's disease. By 1981, 87 of the relatives had developed a dementing illness; in all autopsied cases that illness was diagnosed as Alzheimer's disease.

Notably, all these cases were found among relatives of 51 of the 125 original subjects, implying that about 40 percent of those original patients had a familial disease. Perhaps they carried a gene causing a variant of the disease.



**CEREBRAL BLOOD FLOW diminishes in Alzheimer's disease, as is shown when images of a normal brain (*top*) and of the brain of a patient with the disease (*bottom*) are compared. The false-color images depict a horizontal brain section at eye level. Yellow and red areas are those with high blood circulation. The flow of blood to the temporo-parietal cortex (*left*), a region involved in higher brain functions, is seen to be markedly reduced in the patient with Alzheimer's disease. The images were made by B. Leonard Holman and Thomas C. Hill of the Harvard Medical School, who exploited a technique called single-photon emission tomography. A molecule labeled with a radioactive isotope of iodine is injected intravenously. Single gamma-ray photons emitted by the iodine are detected by probes placed around the patient's head and the signals are processed by a computer to yield the image.**

Perhaps Alzheimer's disease is governed by several genes; in that case a gene carried by the 51 patients could simply increase the likelihood that another and more widespread gene responsible for a nonfamilial variant will be expressed. In another study J. C. S. Breitner and Marshal F. Folstein of the Johns Hopkins University School of Medicine examined the role of genetic factors in cases they classified as being either mild or severe. Analysis of their data led them to conclude that there are two distinct forms of Alzheimer's disease: a relatively mild nonfamilial form generally affecting very old people and a genetically transmitted variant that accounts for about three-fourths of all cases. They calculate that for someone with an afflicted parent or sibling the risk of developing the disease by the age of 80 is about 17 percent.

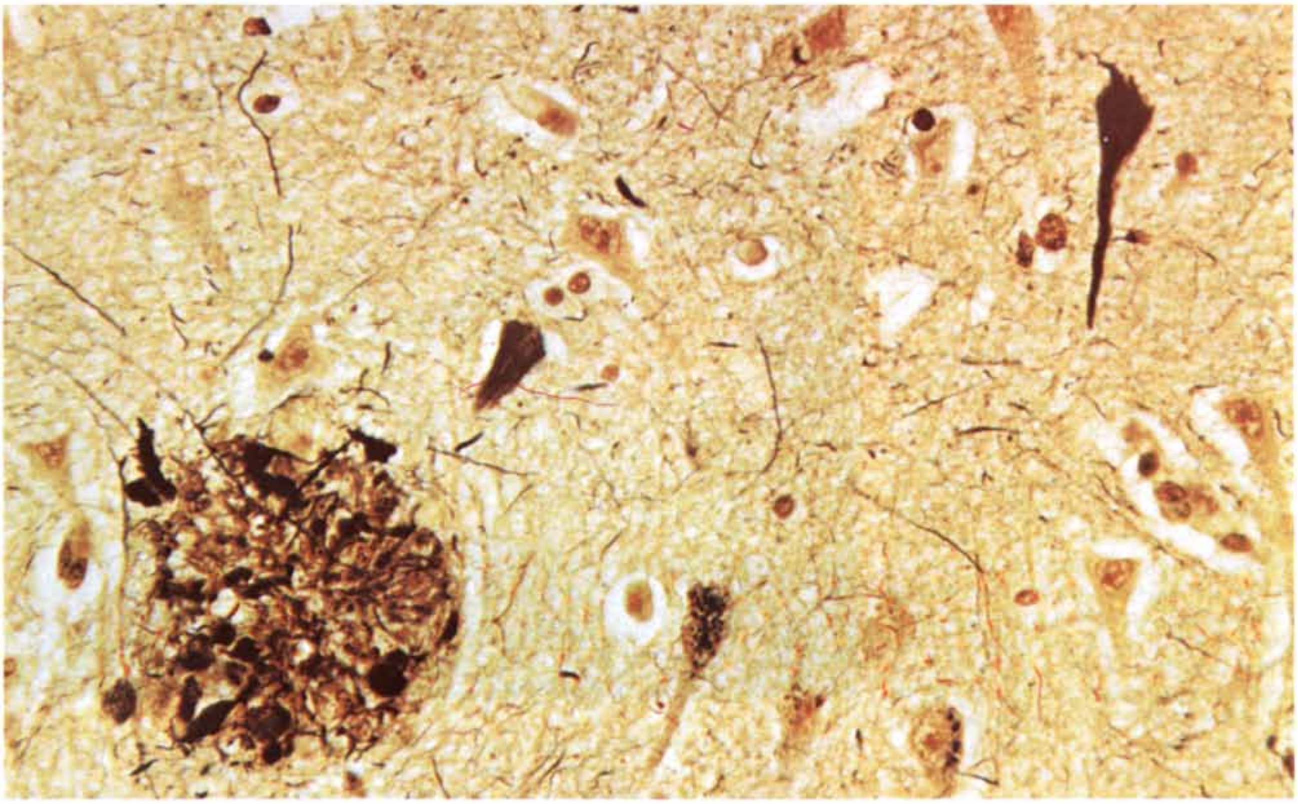
Aside from the epidemiological evidence, the genetic model is supported by the fact that nearly all people with Down's syndrome, a known genetic disorder, develop Alzheimer's disease by about the age of 40. An argument against the model, on the other hand, is that no one has yet identified either a functionally abnormal protein or an abnormal chromosome in Alzheimer's disease.

## The Abnormal-Protein Model

Whatever else it is, Alzheimer's disease is clearly associated with abnormal protein structures. Its three major pathologic signs are the neurofibrillary tangles within neurons, the amyloid that surrounds and invades cerebral blood vessels and the amyloid-rich plaques that replace degenerating nerve terminals. Each of these signs reflects an accumulation of proteins not normally found in the brain. Are the proteins of neurofibrillary tangles and amyloid just unusual accumulations of normal proteins? Or are they qualitatively abnormal, having sequences of amino acids (the subunits of proteins) that are different from those found in normal brain proteins?

If the proteins are abnormal, what does that say about their origin? Is their synthesis directed by abnormal genes? Alternatively, are precursor proteins encoded by normal genes and then subjected to abnormal modification, perhaps by an abnormal enzyme or by one that is improperly activated (possibly by an environmental toxin)? Do the proteins give rise to the observed loss of neurons, either because they are toxic to neurons or because they crowd neurons physically? Could the proteins actually be infectious agents that cause the disease? Or are the protein accumulations unrelated to whatever process destroys neurons and present only as an indication that the destruction has taken place? Only a few of these questions even begin to have answers. Some may emerge from close study of the proteins themselves.

The bundles of fibrous proteins known as neurofibrillary tangles are found in neuronal cell bodies in the hippocampus and the cerebral cortex. They are also particularly abundant in brainstem neurons that release transmitters (notably acetylcholine) found to be deficient in Alzheimer's disease. As Alzheimer noted in 1907, the tangles take up silver stains, which render them visible with the light microscope, at first as dark bands in the cytoplasm of the cell and later as large, fibrous masses that distort the cell body. They are resistant to chemical or enzymatic breakdown and are insoluble in water,



**ABNORMAL PROTEIN STRUCTURES seen at autopsy clearly establish a diagnosis of Alzheimer's disease. Two of these neuropathologic signs, a neuritic plaque (or senile plaque) and neurofibrillary tangles, are seen in a photomicrograph made by Daniel P. Perl of the University of Vermont College of Medicine. A section of the hippocampus was treated with a silver stain and is enlarged some 200 diameters. The large circular structure at the lower left is a neuritic plaque. It consists largely of a collection of degenerating nerve fibers, probably intermixed with aggregates of the abnormal proteins collectively called amyloid. The scattered, roughly triangular dark objects are neurons whose cytoplasm is filled with neurofibrillary tangles, which are bundles of abnormal amyloid fibrils.**

and so they survive in brain tissue long after the neuron in which they arose has died and disappeared.

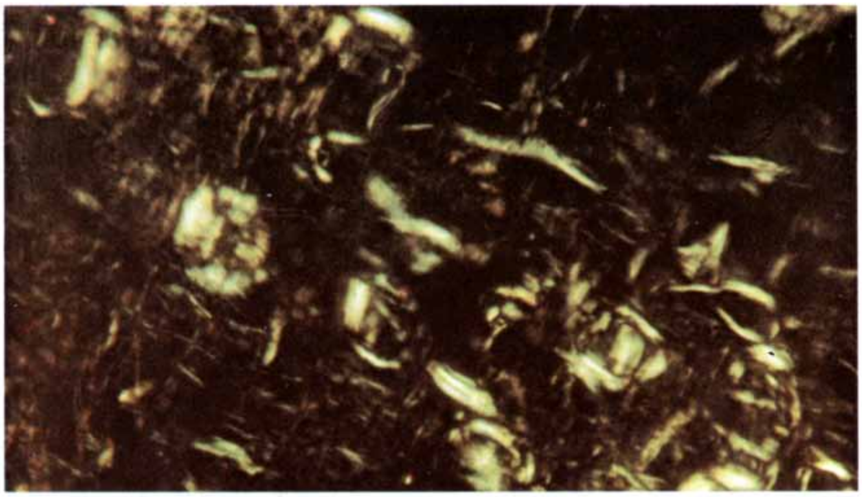
Electron micrographs reveal that most of the individual fibers seen in a photomicrograph of a tangle actually consist of two filaments wrapped around each other to form a helix; they are called paired helical filaments, or PHF. Normal neurons contain three kinds of fibrous proteins with structural and transport roles in the cell: microtubules, microfilaments and neurofilaments (which are also called intermediate filaments). Studies by Dennis J. Selkoe of the Harvard Medical School and by William W. Schlaepfer of the University of Pennsylvania School of Medicine suggest that PHF may be altered neurofilaments. The amino acid composition of purified neurofilaments is similar in many respects to that of PHF, and monoclonal antibodies that bind neurofilament proteins also bind some PHF proteins; the lack of identity could be accounted for by the fact that PHF samples have so far been only partially purified.

The term amyloid, or starchlike, is applied to pathological accumulations within tissues of a protein-rich mass notable mainly for its staining properties: when amyloid is stained with a dye called Congo red and viewed under polarized light, it emits a greenish yellow glow. Some amyloid is seen in the brain of most old people and in other organs, such as the liver and kidney, of people with certain chronic diseases. Abundant cerebral amyloid is, however, almost diagnostic of Alzheimer's disease, where it is seen in the two forms I have mentioned: as deposits in and adjacent to blood vessels and as a component of neuritic plaques. (The abnormal proteins of the neurofibrillary tangles also can exhibit the staining properties of amyloid.) The nature of the protein in amyloid can vary, depending apparently on what precursor proteins are available in a particular disease state.

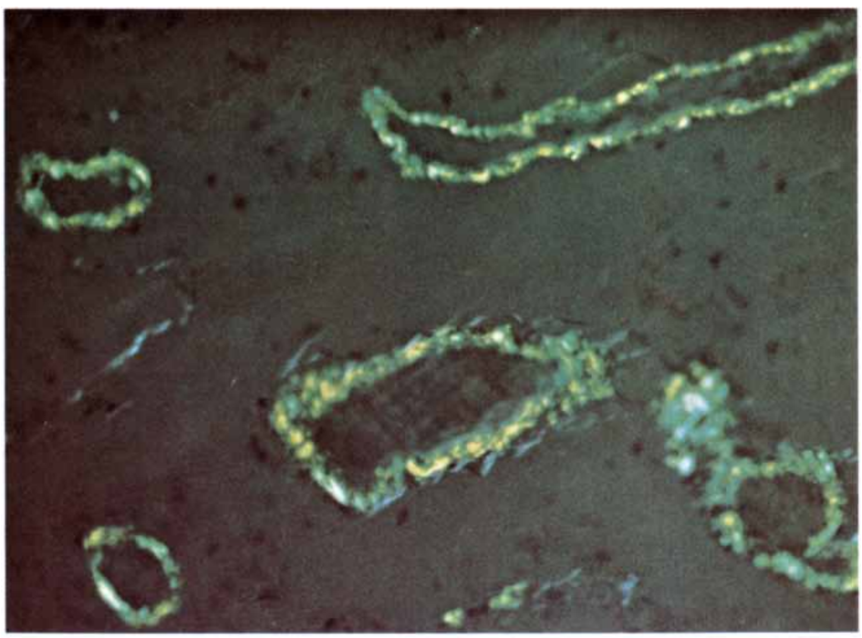
Cerebrovascular amyloid is the commonest finding in Alzheimer's disease. George G. Glenner of the University of California at San Diego School of Medicine examined a series of 350 brains of Alzheimer's disease patients and found the deposits in 92 percent of them. The amyloid usually begins to collect in the muscular middle layer of a brain blood vessel and advances outward; sometimes it entirely replaces the vessel wall, perhaps weakening it and leading to hemorrhage. This process may explain the cerebrovascular hemorrhages that sometimes occur late in the course of Alzheimer's disease, which can make it hard to tell whether a patient is suffering from the disease or from the dementia that can follow multiple small strokes.

Glenner has proposed a theory for the pathogenesis of Alzheimer's disease based on cerebrovascular amyloid. He suggests that "amyloidogenic" proteins in the blood of a patient may be converted, by an enzyme in cerebral blood vessels, into amyloid; the amyloid damages the vessels, causing the leakage into brain tissue of other blood proteins; these proteins are toxic to neurons and activate an enzyme that converts neurofilaments into paired helical filaments, which further damage the neurons; eventually the damaged neurons, with their neurofibrillary tangles, are engulfed in neuritic plaques.

The neuritic (or senile) plaque is the third member of the triad whose presence signals Alzheimer's disease to the neuropathologist. Plaques are usually most abundant in the cerebral cortex and hippocampus and in the amygdala, a nucleus of cells near the hippocampus that seems to be particularly



**AMYLOID FIBRILS** of neurofibrillary tangles appear as bright greenish yellow flame-shaped structures in a polarization photomicrograph made by Dennis J. Selkoe of the Harvard Medical School. Brain tissue was fractionated by density-gradient centrifugation. A fraction highly enriched in the tangles was stained with the dye Congo red. The amyloid is birefringent: it reflects polarized light primarily at the greenish yellow wavelength.



**CEREBROVASCULAR AMYLOID** is the third major sign of Alzheimer's disease. A polarized-light micrograph of cerebral blood vessels stained with Congo red, made by George G. Glenner of the University of California at San Diego School of Medicine, shows deposits of birefringent amyloid fibrils in vessel walls. The enlargement is 200 diameters.

damaged in the disease; the number of plaques tends to correlate closely with the extent of dementia. Within each region the plaques are localized in areas containing the axons and terminals of neurons rather than their cell bodies. Henryk M. Wisniewski and Robert D. Terry, then working at the Albert Einstein College of Medicine, observed that the plaques undergo an orderly change in appearance as they mature. At first they are spherical masses, anywhere from 50 to 200 micrometers (thousandths of a millimeter) in diameter, of degenerating nerve terminals, amyloid and the kind of glial cell (the supporting cells of the brain) associated with inflammatory reactions. Then the amyloid accumulates within a central core, surrounded by degenerating axons, macrophages (scavenger cells) and more glia. Ultimately the cellular debris disappears, leaving discrete deposits of amyloid.

In addition to accumulating these three kinds of abnormal proteins, the brain of Alzheimer's disease patients seems to synthesize subnormal quantities of proteins in general. Charles A. Marotta of the Harvard Medical School and his colleagues think the reason may be that the brain is deficient in RNA (the nucleic acid that mediates the translation of DNA to manufacture protein) because an enzyme that breaks down RNA is improperly regulated.

## The Infectious-Agent Model

If patients with Alzheimer's disease often look well, and if they never show such signs of a brain infection as fever or the presence of white blood cells and proteins in the cerebrospinal fluid, why should one consider the possibility of an infectious origin? There are two reasons: scrapie and Creutzfeldt-Jakob disease. The first is a slowly progressive, invariably fatal brain disease that is widespread in sheep and goats. The second is a rather rare brain disease (one new case is discovered per million of population per year) that usually affects people between 55 and 75; it causes progressive dementia, disturbances in posture, vision and the control of movement, and death comes about a year after the onset of the disease.

The finding some years ago that both diseases can be transmitted to animals (by the injection into the brain of extracts of infected tissue) indicated they must be caused by an infectious agent, presumably a virus; the causative agent can be isolated from the brain and from other tissues. Yet neither disease gives rise to any evidence of brain inflammation or of functional disturbance in any other organ. Moreover, in both diseases a remarkably long incubation period (roughly a fourth of the affected individual's normal life span) may pass between exposure to the agent and the first appearance of clinical signs. In both cases the putative infectious agent has therefore been thought to be an unconventional, "slow" virus.

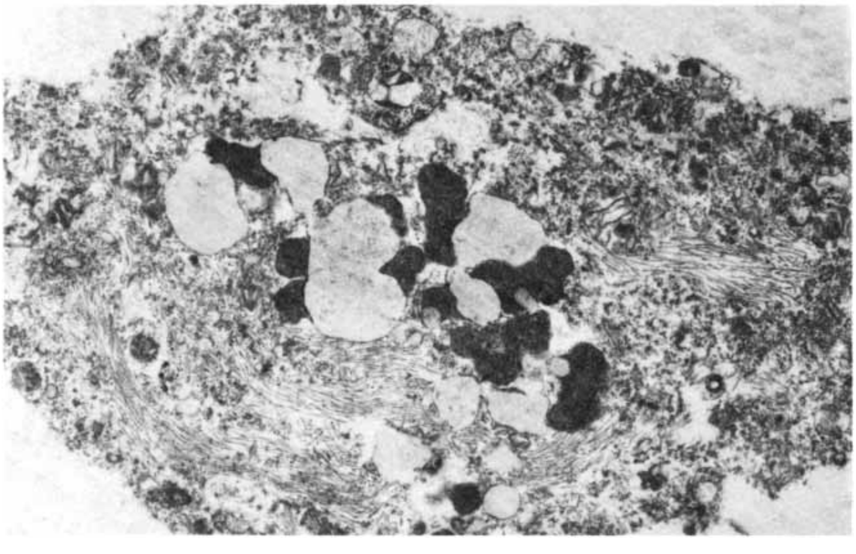
Now it appears that the agents causing scrapie and Creutzfeldt-Jakob disease may not be viruses at all but an even more unconventional kind of agent called a prion: a protein particle first isolated from the brain of sheep with scrapie by Stanley B. Prusiner of the University of California at San Francisco School of Medicine. Unlike any other known infectious agent, prions apparently contain neither DNA nor RNA. Patricia A. Merz and R. A. Summerville of the New York State Institute for Basic Research on Developmental Disabilities observed that highly purified preparations of the scrapie agent contain numerous rodlike structures, which they suggested might be the infectious agent itself. Prusiner thinks the rods are aggregates of prions. When the rods are stained with Congo red, they look much like the amyloid seen in the brain of sheep with scrapie and of some people with Creutzfeldt-Jakob disease. Prusiner suggests this amyloid may actually be an aggregate of prion rods. He thinks that the amyloid of Alzheimer's disease may also be a collection of prions and that therefore prions may be the agents of Alzheimer's disease as well.

D. Carleton Gajdusek of the National Institute of Neurological and Communicative Disorders and Stroke and his colleagues have observed an abnormal structure made up of twisted filaments in the brain of animals with scrapie and people with Creutzfeldt-Jakob disease or kuru. (Kuru is a neurological disorder affecting certain tribes in New Guinea. It was for demonstrating the transmissibility of kuru and Creutzfeldt-Jakob disease that Gajdusek was awarded a Nobel prize in 1976.) They have proposed that this filamentous structure may be the infectious agent in all three disorders. The structures are not present in Alzheimer's disease, but they do share some immunological properties with the amyloid fibrils of neurofibrillary tangles.

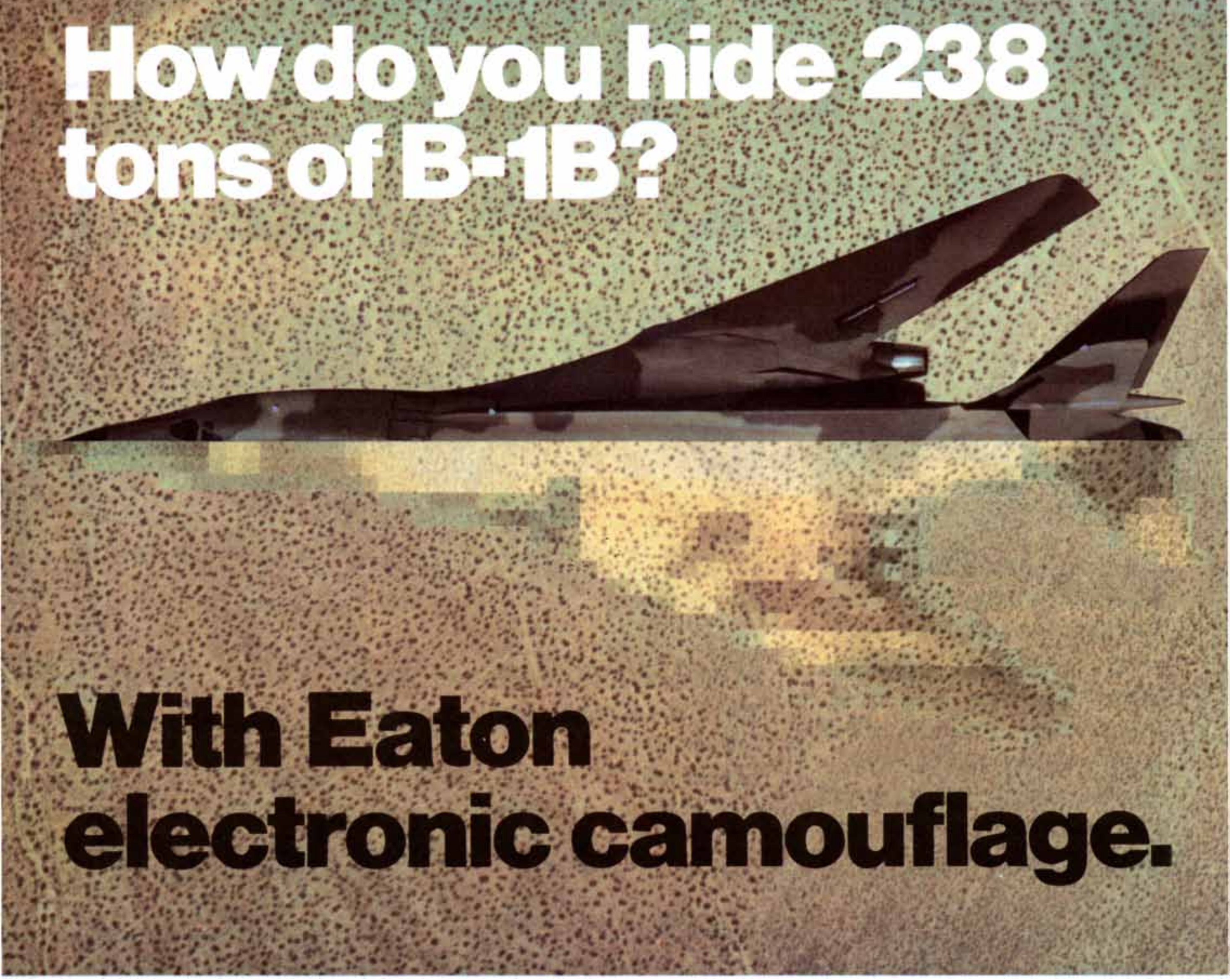
If Alzheimer's disease is caused by an infectious agent, it should be transmissible, but attempts to transmit it to experimental animals have been unsuccessful. Perhaps infection by the agent of the disease (if there is one) requires a particular genetic makeup, a concurrent immune disorder or prior exposure to an environmental toxin.

## The Toxin Model

Some investigators believe there is evidence that salts of aluminum may contribute to the development of Alzheimer's disease. Such salts are present in drinking water, may be added to foods and drugs (including some processed cheeses, antacids and buffered aspirins) and may be released from aluminum cans and utensils. Daniel P. Perl of the University of Vermont Col-



**NEUROFIBRILLARY TANGLES within a nerve cell are resolved as individual fibers in this electron micrograph, made by Selkoe. A thin section of a cortical neuron isolated from the brain of a patient is enlarged about 10,000 diameters. The fibers course through the cytoplasm around an aggregation of the fatty pigmented granules called lipofucsin.**




Four associate prime contractors created the world's most sophisticated airplane, with the U.S. Air Force as overall systems manager. Rockwell International builds the airframe, General Electric builds the engines, Boeing provides the offensive avionics. And Eaton's task is to conceal the whole thing from the enemy—all 238 tons of it.

So Eaton created a unique defensive electronics system that can detect and attach priorities to hostile threats anywhere in the world, and direct appropriate jamming responses—in milliseconds.

Why Eaton? Because in certain areas of electronics, we are acknowledged leaders. That's why we build the microwave landing system for the space shuttle. And why we supply more air traffic control processing and display systems than any other company in the world.

You may know Eaton best as a major supplier of state-of-the-art components for the automotive and industrial sectors. Which we are. But Eaton is also a major participant in the exciting business of high-technology electronics. Eaton Corporation, Eaton Center, Cleveland, Ohio 44114.

EATON

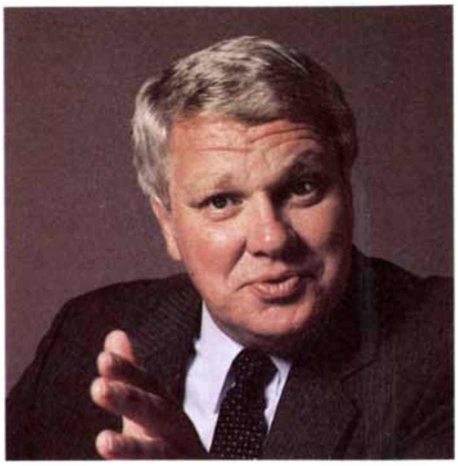
# WHAT SHOULD YOU
# WORKING "PARTNERSHIP"

## RESULTS WITH ASSURANCE.

**Deeds more than words have characterized our working partnerships with customers large and small. We feel the best assurances we can offer prospective customers are the productive experiences and results of our present customers.**

**"At Container Corporation of America, an HP computer network helps us respond to customers faster, and saves us $600,000 a year."**

*Jeff Norkin*
*Vice President*



**"At Thomas & Betts Corporation, six HP computer-aided test and data acquisition systems paid for themselves in less than six months."**

*Don Michel, Manager*
*Insulation Displacing Connector Engineering*

### Your long-term needs addressed.

HP's working partnership with you begins before the initial sale. Our objective is to meet your needs over the long term, and to expand the relationship wherever the "fit" is good. If your needs cannot be met with an HP-only solution, we are not timid about working with other organizations to achieve the best possible results for you.

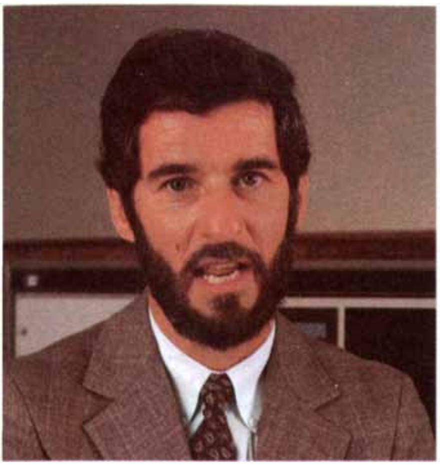
### More than computers.

Beyond computers, HP offers a great variety of instrumentation and computer-based systems for productive solutions in many areas: office information systems; computer-aided design, test, and manufacturing; laboratory automation; patient monitoring and hospital financial management—to name a few.

### Quality beyond products.

Quality at HP includes all aspects of our relationship with customers. The same quality and reliability engineered into HP products is built into our customer support, training, and service programs. Your expectations of HP quality should be systems operating effectively at least 99 percent of the time.

# EXPECT FROM A WITH HEWLETT-PACKARD?



**'At City National Bank, four HP computers deliver mainframe capability at one-eighth the cost."**

*Newton Gore*
*Sr. Vice President*

**"At Deaconess Medical Center, HP CARE-NET systems improve patient monitoring and maintain close to 100% reliability."**

*Bim Lindsey*
*Manager, Cardiovascular Services*

**"At American Cyanamid, three HP lab automation systems speed data reduction and help handle a twofold increase in assays."**

*Dr. William Hodgson*
*Director of Scientific Services*

## R&D that protects your investment.

The long-term cost of software usually exceeds the cost of hardware several times over. Much of HP's R&D concentrates on technology that will preserve your software investment while increasing the performance and decreasing the cost of the hardware. Doing business with HP is a hedge against obsolescence.

## Satisfaction. Based on 45 years of meeting exacting customer expectations.

From the outset, HP has built its business by satisfying the high performance expectations of scientists and engineers. They can't accept marginal performance, uncertain quality, or spotty support. Against this background of unequivocal performance HP builds working partnerships with all of its customers, business and technical.

In summary, **working partnership is inherent in our name—** and fundamental in our dealings.

HEWLETT PACKARD

0002409A

# BMDPC STATISTICAL SOFTWARE
## CONFIDENCE AND CONVENIENCE

**BMDP® is Data Analysis with Confidence.** Researchers and statisticians know quality research demands quality analysis. That's why they rely on BMDP Statistical Software for all their data analysis needs.

BMDP is a comprehensive library of dependable statistical programs. You can use BMDP to perform many types of analysis. Capabilities of BMDP range from simple data display and description to the most advanced statistical techniques. In fact, BMDP provides the widest range of statistical approaches required to effectively analyze your research data:

- **Data Description**
- **Data in Groups**
- **Plots and Histograms**
- **Frequency Tables**
- **Missing Values**
- **Nonlinear Regression**
- **Analysis of Variance and Covariance**
- **Regression**
- **Multivariate Analysis**
- **Nonparametric Analysis**
- **Cluster Analysis**
- **Survival Analysis**
- **Time Series**

Powerful statistical and computing procedures make BMDP especially appealing to the experienced data analyst. However, you need not be a seasoned statistician or computer expert to perform data analysis with confidence. English-based instructions make BMDP easy to learn and easy to use.

As the acknowledged leader in statistical methodology, BMDP has become the standard by which others measure their performance. The BMDP programs are known for superior technology and state-of-the-art methodology. BMDP is committed to developing new statistical methods for data analysis, and continues to lead the way in providing superior statistical software.

**BMDP is Data Analysis with Convenience.** With our latest advance, you can even perform complex statistical analyses without a large computer. BMDPC offers you both the confidence of BMDP and the convenience of using your own PC.

The same programs that have been trusted for more than 20 years by researchers worldwide are now available for the IBM Personal Computer and compatibles.

Best of all, the BMDPC programs are available in low-cost subsets. So, you can pick and choose programs for your PC. Pick the types of analyses you use most, and choose from our broad selection of statistical programs.

Put the statistical software of choice on your IBM® PC*, so you can perform data analysis with confidence and convenience.

### Hardware Requirements

Minimum requirements for running BMDP on the PC include the following:

- a 5 megabyte hard disk
- the capacity for double-sided, double density floppy diskettes
- an 8087 floating point processor
- a PC/DOS 2.0 or later operating system
- 640 kilobytes of memory.

### FOR MORE INFORMATION WRITE OR CALL

**BMDP Statistical Software**
1964 Westwood Blvd., Suite 202
Los Angeles, California 90025
Phone (213) 475-5700

BMDP is a registered trademark of BMDP Statistical Software

IBM PC® is a registered trademark of International Business Machines lege of Medicine finds that aluminum accumulates preferentially in human neurons harboring neurofibrillary tangles. Igor H. Klatzo and Wisniewski, then working at the National Institute of Neurological Diseases and Stroke, reported that the injection of aluminum salts into rabbits or cats (but not mice, rats or monkeys) generates neurofibrillary tangles. The tangles appear when the concentration of aluminum in the brain reaches a level that can be attained in Alzheimer's disease, Donald C. Mclaughlin of the University of Toronto Faculty of Medicine has shown.

According to Allen Alfrey of the University of Colorado School of Medicine, an irreversible dementia seen in some people who have undergone repeated kidney dialyses with aluminum-rich dialysis solutions is associated with elevated brain aluminum levels. It is known that some brain enzymes can be inhibited by aluminum and that aluminum salts can impair the transport of certain proteins (including precursors of neurofilaments) from a neuron's cell body down the axon to the nerve terminal. Gajdusek has found high brain levels of aluminum to be associated with a dementia seen in natives of certain communities in Guam; a lack of adequate calcium in their drinking water seems to enhance their absorption of aluminum.

The evidence implicating aluminum can be challenged. The association of aluminum with neurofibrillary tangles may mean only that the tangles, once formed, happen to have an affinity for aluminum. The particular neurofibrillary tangles found to be induced in rabbits and cats have single-strand filaments rather than paired helical filaments, and they appear in parts of the nervous system that are not affected in Alzheimer's disease. Some signs and symptoms of the dialysis-associated dementia are not present in Alzheimer's disease. It is possible that aluminum cannot by itself give rise to the clinical and pathologic indicators of the disease but that its presence contributes to their appearance in people exposed to another causative factor.

**PAIRED HELICAL FILAMENTS are resolved when the fibers of a neurofibrillary tangle are further enlarged (to some 110,000 diameters in this electron micrograph, also made by Selkoe). Each fiber is seen to be composed of two filaments, which are twisted into a helix.**

## The Blood-Flow Model

There is irony in the fact that blood flow is being considered as a factor in Alzheimer's disease. Not long ago most adult dementia was blamed on "poor circulation" or "hardening of the arteries." It then became clear that such dementia is usually correlated not with evidence of atherosclerosis or small strokes but with the acetylcholine deficit and the tangles, plaques and cerebrovascular amyloid of Alzheimer's disease. Now there is a perception that the disease is indeed at least associated with a profound reduction in the amount of blood delivered to the brain, in the amount of oxygen and glucose extracted from this blood and in the energy generated from the oxygen and glucose. The perception arises largely from studies of brain metabolism done with new noninvasive imaging techniques. In positron-emission tomography (PET), for example, a chemical of interest is treated so that it emits positrons; when it is administered to a patient, its distribution in the brain can be assessed by computer analysis of data from an array of positron scanners.

Richard S. J. Frackowiak and J. M. Gibbs of the National Hospital for Nervous Diseases in London have observed that in normal people blood flow to the brain declines by about 23 percent between the ages of 33 and 61; the brain compensates for this deficit by extracting a larger proportion of the oxygen carried by the blood. In Alzheimer's disease there is a further decrease in cerebral circulation but no additional compensatory increase in oxygen extraction; both blood flow and oxygen consumption decline to a level some 30 percent below that seen in nondemented elderly people. Blood flow and oxygen consumption continue to decline as the clinical situation deteriorates. The greatest decreases are detected within the frontal and parietal lobes of the cerebral cortex—which also tend to show severe pathologic changes.

There is a parallel decline in the rate at which the brain consumes glucose, its major source of energy. D. Frank Benson of the University of California at Los Angeles School of Medicine found that in the brain of patients with Alzheimer's disease from 30 to 50 percent less glucose was consumed in each of the four cortical regions and one subcortical region he could visualize; in patients whose dementia resulted from multiple small strokes, on the other hand, the glucose-consumption deficit tended to be confined to the regions affected by the strokes. John P. Blass of the Cornell University Medical College has observed that phosphofructokinase, an enzyme necessary to convert glucose to high-energy intermediates, is deficient in brain samples and in cultured skin cells from patients with Alzheimer's disease.

Are the reductions in the brain's ability to obtain chemical energy a cause or a consequence of the disease's damage to neurons? It is true that a decrease in the number of cells consuming oxygen and glucose would be expected to diminish the brain's demand for blood. The observed reduction in blood flow seems often to be greater, however, than what can be accounted for by the loss of brain tissue.

How might the blood-flow disturbances seen in Alzheimer's disease come about? Arnold B. Scheibel of the U.C.L.A. School of Medicine found that small neurons controlling the dilation and contraction of brain arterioles tend to disappear with normal aging and are particularly deficient in the brain of demented patients. Several years ago my colleagues and I at the Massachusetts Institute of Technology found that the release of the neurotransmitter dopamine can control the flow of blood to a region of the rat brain rich in dopamine; others have found that blood vessels have specific receptors for a number of different transmitters, suggesting that the release of these transmitters may affect regional brain blood flow. Perhaps the loss of nerves controlling blood flow prevents the delivery of enough blood to neurons.

### The Acetylcholine Model

In 1976 two groups, led by Peter Davies of the University of Edinburgh Faculty of Medicine and by David Bowen of the Institute of Neurology in London, reported the first clear biochemical abnormality associated with Alzheimer's disease. In the hippocampus and the cerebral cortex of patients, they found, the level of the enzyme choline acetyltransferase (CAT) can be reduced by as much as 90 percent. The enzyme catalyzes the synthesis of acetylcholine from its precursors, choline and acetyl coenzyme *A;* the loss of CAT activity reflects the loss of cholinergic, or acetylcholine-releasing, nerve terminals in these two regions of the brain. Marek-Marsel Mesulam of the Harvard Medical School and Joseph T. Coyle of the Johns Hopkins University School of Medicine found that the missing terminals are those of neurons whose cell body is far away in the basal forebrain or septum.

The dramatic biochemical disturbance identified by Davies and Bowen has been confirmed repeatedly by other investigators, and it seems to many of us to be the clue most likely eventually to point to the cause of Alzheimer's disease. It also suggests an explanation for the disease's cardinal symptom: loss of memory. If CAT levels are reduced in the hippocampus, the level of acetylcholine (which cannot be measured at autopsy) must in all probability also be reduced there. Abundant evidence suggests that cholinergic terminals in the hippocampus are critically important for memory formation. It is therefore plausible to hypothesize that some of the cognitive deficits of Alzheimer's disease are the direct result of a reduction in the acetylcholine-mediated transmission of nerve impulses.

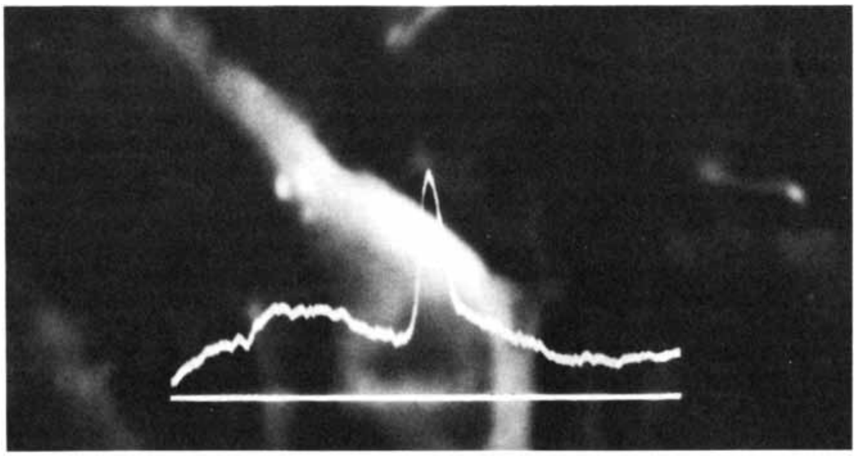
The cholinergic hypothesis suggests an objective for therapy: a drug restoring the acetylcholine level might be effective, much as L-dopa is effective in correcting the deficit in neurotransmission caused by the loss of a different transmitter, dopamine, in Parkinson's disease.

It has not been easy, however, to find a safe and effective agent that will fully restore the acetylcholine level in Alzheimer's disease or that will mimic acetylcholine's activity. The first kind of agent might be an acetylcholine precursor that is converted into the transmitter; the difficulty is that the terminals containing CAT (which would catalyze the conversion) are precisely the ones that are deficient. The acetylcholine level might, on the other hand, be restored by a drug that inhibits the breakdown of acetylcholine; the trouble is that such a drug would probably act not only in the hippocampus and cerebral cortex but also in parts of the brain (and in other organs) where such inhibition would be unacceptable. The same difficulty would arise in the case of a mimicking agent.

To understand how a drug might be developed that is able to restore deficient cholinergic transmission, and to act only where such restoration is wanted, one needs to consider the life cycle of an acetylcholine molecule. The transmitter is manufactured primarily in the terminals from which it is released. The first step in its synthesis is the uptake of choline from the synaptic cleft, the minute space between the terminal of a presynaptic neuron (where the transmitter is released on the arrival of a nerve impulse) and the surface of the postsynaptic neuron (where the transmitter produces its effect by altering the electrical activity of the receiving neuron). The choline can be taken up into the terminal by either of two transport proteins in the presynaptic cell membrane. One of them is present not only on cholinergic neurons but also on all cells, since choline is required for the synthesis of phosphatidylcholine, or lecithin, a constituent of all cell membranes. The other transport protein is found only on cholinergic terminals, where its particularly high affinity for choline enables the terminal to take up almost all the choline available to it in the synaptic cleft.

The next step, catalyzed by CAT, combines the choline with acetyl coenzyme *A* to make acetylcholine. Terminals have an excess of CAT and too little of the precursors choline and acetyl coenzyme *A* to keep the enzyme molecules fully occupied; the rate at which the transmitter is synthesized is therefore limited not by the amount of CAT enzyme but by the level of the two precursors. Once synthesized, acetylcholine molecules are stored in the terminal until the arrival of a nerve impulse discharges some of them into the synapse. Once it is in the synapse, an acetylcholine molecule can cross the cleft and interact with a receptor on the postsynaptic neuron, thereby transmitting the signal generated by the nerve impulse. It can be broken down by the enzyme acetylcholinesterase to generate choline. Or it can interact with a receptor on the presynaptic membrane whence it came, thereby delivering a signal that modulates the subsequent release of acetylcholine from the presynaptic terminal.

Each of these steps in the life cycle of an acetylcholine molecule provides the neuropharmacologist with opportunities to find drugs that either mimic the deficient transmitter's postsynaptic action or increase the amount of the



**CONCENTRATION OF ALUMINUM is revealed in a scanning electron micrograph, made by Perl, of a neuron from the brain of an Alzheimer's disease patient. The brain sample was treated with a silver stain, which accumulated in a neurofibrillary tangle that fills the cytoplasm of one cell. The image was formed first, by secondary electrons emitted by the sample when an electron beam was scanned over it. Then the beam was swept across the cell once again (*white line*) and X rays emitted at a wavelength characteristic of aluminum were collected (*white curve*), showing an aluminum peak over the cell nucleus.**

transmitter, released from remaining cholinergic terminals, that is present in synapses. One approach would be to find a drug that simulates acetylcholine's postsynaptic effect, and does so only in regions where the transmitter is deficient. Some very preliminary results reported by Robert E. Harbaugh of the Dartmouth-Hitchcock Medical Center suggest this may be possible.

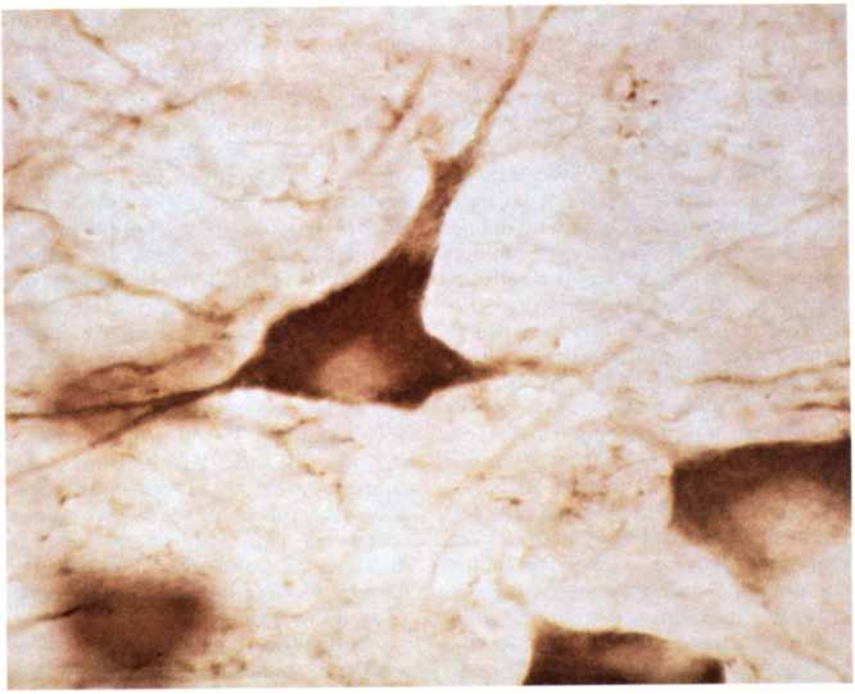
An alternative approach is to increase the release of acetylcholine. One way to do it would be to find a drug that selectively blocks the presynaptic acetylcholine receptors on surviving cholinergic terminals. In the absence of a modulating signal from those receptors, the terminal should release more acetylcholine per firing. Or a drug might be found that prolongs the activity of whatever acetylcholine is available by blocking the breakdown enzyme, acetylcholinesterase, only where the transmitter is in short supply.

Perhaps the most obvious strategy is to enhance acetylcholine synthesis in surviving terminals by providing the key enzyme, CAT, with more choline. It happens that CAT is more responsive to additional choline when the neuron containing it is firing often than it is when the neuron is quiescent. Cholinergic neurons tend to fire more often in regions where acetylcholine release is deficient. Therefore one can hope that supplemental choline given to a patient would have a selective effect in acetylcholine-deficient regions: precisely where the effect is wanted.

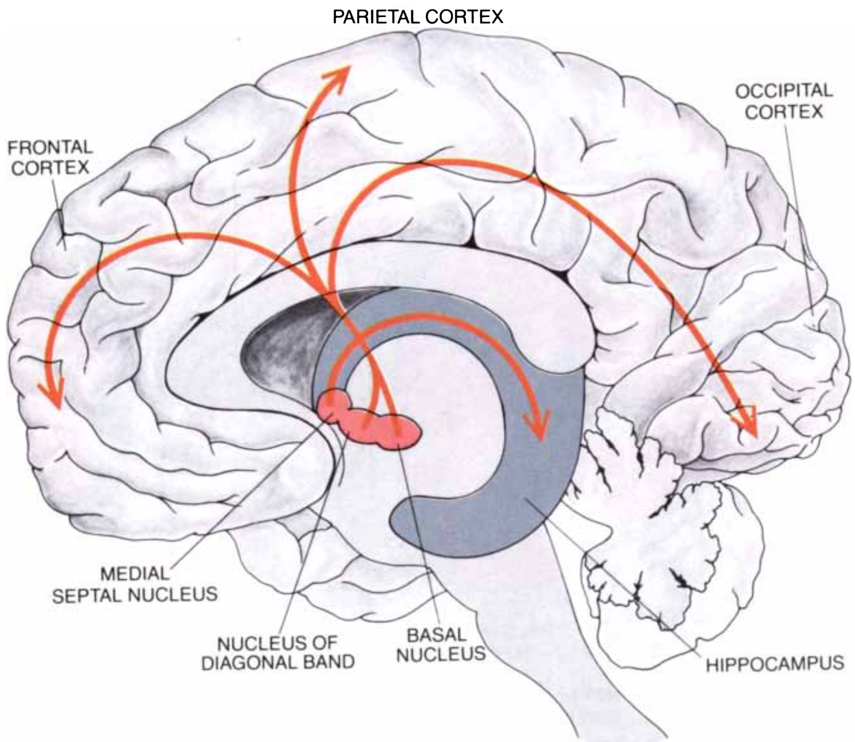
In my laboratory we have given healthy volunteers supplemental choline (or lecithin, which the body can break down to yield choline). The choline concentration in their blood rose to levels that in rats are high enough to increase the choline level in the brain and thereby step up the release of acetylcholine from neurons that are firing frequently. John H. Growdon of the Harvard Medical School and M.I.T. found this treatment also increased the choline level in the subjects' cerebrospinal fluid, indicating that the extra choline does have access to the brain.

Initial reports in 1978 and 1979 describing cognitive improvement in a few patients treated for short periods (usually with impure lecithin) failed to be confirmed when patients were treated for a few weeks under controlled conditions. Now the effort is to see whether prolonged treatment (for periods during which untreated control patients show measurable deterioration) improves behaviors impaired in Alzheimer's disease or at least slows deterioration.

The only long-term, carefully con-



**CHOLINERGIC TERMINALS, which synthesize the neurotransmitter acetylcholine and therefore contain the enzyme choline acetyltransferase (CAT), are identified in a photomicrograph made by Marek-Marsel Mesulam of the Harvard Medical School. To distinguish these cells a section from the basal nucleus of a rhesus monkey brain was treated with a monoclonal antibody binding only to CAT. The enzyme horseradish peroxidase, linked to the antibody, was subjected to a reaction whose product stained cholinergic cells brown.**



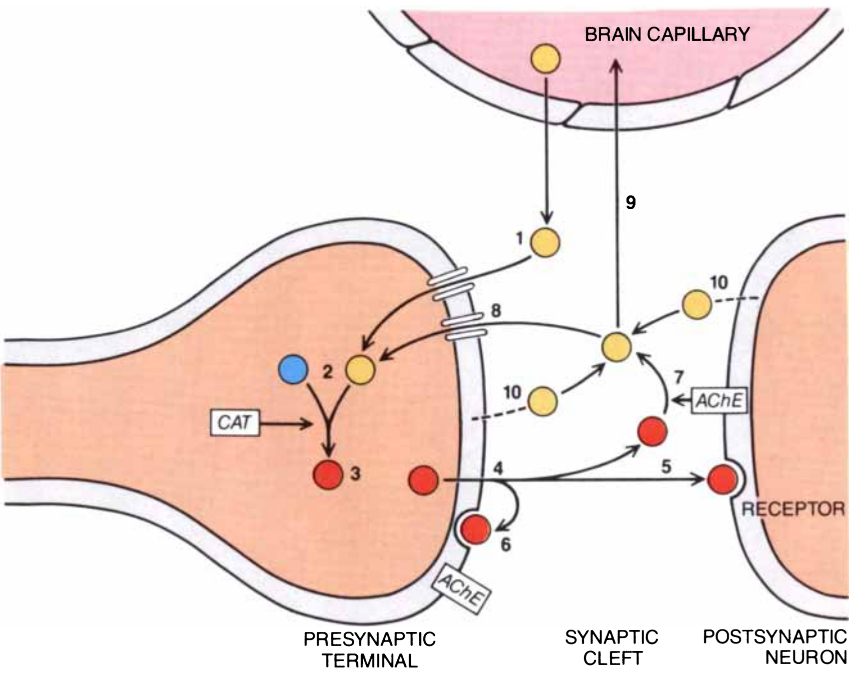

**MAJOR CHOLINERGIC PATHWAYS implicated in Alzheimer's disease are indicated in this diagram based on the findings of Mesulam and Joseph T. Coyle of the Johns Hopkins University School of Medicine and his colleagues. There is a deficiency of CAT, and therefore presumably of acetylcholine, in the cortex and hippocampus of patients with Alzheimer's disease. The nerve terminals that would normally contain CAT belong to cholinergic neurons whose cell bodies are in one of the indicated nuclei at the base of the brain.**

trolled study of the effect of supplemental purified lecithin has recently been completed. Raymond Levy of the University of London Faculty of Medicine found continuing behavioral improvement in eight out of 24 Alzheimer's disease patients treated. The average age of those eight was 79, whereas the average age of those who did not improve was 69. Were the responders patients whose brain lesions were incomplete and perhaps limited to cholinergic neurons? Various lines of evidence suggest that there is more than one form of Alzheimer's disease, and specifically that people who develop the disease late in life tend to have a milder form whose effect is largely confined to cholinergic neurons.

The selective reduction in brain acetylcholine observed in Alzheimer's disease may provide a clue for understanding the pathogenesis of the disease: the mechanisms whereby some causative factor gives rise to the changes in behavior and in brain tissue. In particular, what kind of process could make the long-axon cholinergic neurons leading from the base of the brain to the hippocampus and cerebral cortex particularly vulnerable?

One mechanism could be a kind of autocannibalism of neuronal membranes. As I have mentioned, cholinergic neurons are the only cells that need choline for two purposes: as a precursor both for a neurotransmitter and for a constituent (phosphatidylcholine) of their cell membrane. They would be particularly vulnerable if choline were in short supply. (The choline level is not depressed in the blood or cerebrospinal fluid in Alzheimer's disease, but it might still be subnormal in some cholinergic terminals if their capacity to take up choline or to synthesize lecithin is diminished or if they consume excessive amounts of choline to make acetylcholine.) Faced with a lack of free choline, the terminals of these neurons might break down the phosphatidylcholine in their own membrane to get the choline they need to make acetylcholine. That could impair membrane function and ultimately destroy the terminals.

Cholinergic neurons could also be uniquely vulnerable because a tropic, or stimulating, factor they need is missing; a search for some such factor is under way in several laboratories. Finally, if the disease process is triggered by an infectious agent, the basis of selective vulnerability could simply be the agent's predilection for long-axon cholinergic neurons.




**METABOLISM AND MOVEMENT of an acetylcholine molecule are traced. Choline (*1*) delivered to the synaptic cleft by the blood enters a presynaptic cholinergic terminal, where CAT combines it with acetyl coenzyme *A* (*2*) to form acetylcholine (*3*). When the arrival of a nerve impulse depolarizes the neuron, acetylcholine is released (*4*). It can cross the synapse to interact with a receptor on the postsynaptic neuron, thereby transmitting the signal generated by the nerve impulse (*5*). It can instead interact with a receptor on the presynaptic membrane (*6*), thereby modulating the further release of acetylcholine. Or it can be broken down by acetylcholinesterase (*AChE*), yielding choline (*7*), which is taken up by the terminal (*8*) or carried off in the blood (*9*). The breakdown of phosphatidylcholine (lecithin), a constituent of cell membranes, contributes to the choline supply (*10*).**

Focusing on acetylcholine has a disadvantage: it obscures the fact that in Alzheimer's disease the brain often exhibits major deficits in other transmitters (particularly norepinephrine and serotonin) as well. In the hippocampus and cerebral cortex these transmitters, like acetylcholine, are found largely in terminals of neurons whose cell body lies in more primitive brain centers. It seems strange that in a disease notable for giving rise to disturbances in higher, cortical functions the known neurochemical changes are largely traceable to subcortical neurons.

Actually, however, a large number of cortical neurons are also lost in Alzheimer's disease, and one relatively minor cortical neurotransmitter, somatostatin, has been shown to be reduced in some patients. A recent report from a group at the University of Iowa College of Medicine suggests that hippocampal neurons mediating the transfer of information between the hippocampus and other regions may also be specific targets. The inability to associate any particular transmitter deficiency with the loss of such neurons could mean that they release an unidentified transmitter or that they share a transmitter with so many other, unaffected neurons that their loss does not detectably reduce the transmitter's level in the brain.

## The Elephant Model

Different investigators tend to see, or at least to focus on, six different sets of manifestations of Alzheimer's disease, and they collect six different sets of data; they then assume, at least for the purpose of investigation, that Alzheimer's disease has its origin in genetics, protein accumulation, infection, a toxin, a neurochemical disturbance or vascular insufficiency. Just so did the six blind men conclude that an elephant was very much like a wall, spear, snake, tree, fan or rope—depending on the part they happened to touch. The elephant was all of these, but they had not perceived its essential elephantness.

Not many years ago different investigators assumed, at least for the purpose of investigation, that cancer is caused by viruses, foods, faulty genes or impaired regulation of genes, environmental toxins, radiation or immune-system deficiencies. All the investigators turned out to be correct; now they are converging on what may be the essential "elephantness" of cancer. In time the same thing may happen for Alzheimer's disease.

# Hot Nuclear Matter

*Ordinarily the nucleus of an atom is rather like a droplet of liquid. High-speed collisions that heat and compress it can give rise to new phases of nuclear matter: a vapor and perhaps a solid and a plasma*

by Walter Greiner and Horst Stöcker

The nucleus of the atom is a bit of matter quite unlike anything familiar in everyday experience. It is extraordinarily dense: a thimbleful of nuclear matter would weigh a million tons. Nevertheless, the substance of the nucleus can be described in terms of the same physical properties that characterize any other form of matter. Of particular interest is the response of nuclear matter to changes in temperature, density and pressure. Water undergoes dramatic changes when the temperature is raised from zero degrees Celsius to 100 degrees. Nuclear matter might be subject to analogous transformations, but up until now there have been no experimental methods for inducing them. It is as if water could be studied only at room temperature and pressure and nothing were known of ice or steam.

What has changed this situation is the development of relativistic heavy-ion accelerators, which bring about high-energy collisions between massive nuclei. The collisions are said to be relativistic because the nuclei move at nearly the speed of light, so that effects of the special theory of relativity become important. Such collisions offer the unique opportunity to compress nuclear matter to several times its normal density and to heat it to temperatures in excess of $10^{12}$ degrees. Comparable conditions exist inside a neutron star; the interior of the sun, in contrast, is at a temperature of only $10^6$ degrees. Still more powerful accelerators may be able to create on a nuclear scale the kinds of matter that existed at the birth of the universe in the big bang.

The accelerators employed in work with relativistic heavy ions are synchrotrons, like many of those used in elementary-particle physics. Instead of accelerating individual particles such as protons or electrons, however, they create a beam of ions: atoms that have been stripped of most of their electrons. The nucleus accounts for almost all of the mass of an ion, and so one can think of it as simply an accelerated atomic nucleus.

The capabilities of a heavy-ion accelerator depend on how much energy it can impart to a nucleus and on how large a nucleus it can accelerate. The energy is generally expressed in millions or billions of electron volts (MeV or GeV) per nucleon; the size of the nucleus is simply the total number of nucleons, equal to the atomic mass number. The synchrophasotron at the Joint Institute for Nuclear Research at Dubna in the U.S.S.R. can accelerate ions of neon (atomic mass 20) to an energy of 4 GeV per nucleon; at that energy the ions move at 98 percent of the speed of light. The Bevalac at the Lawrence Berkeley Laboratory of the University of California reaches 2 GeV per nucleon with ions as heavy as uranium (atomic mass 238).

The accelerated nuclei collide with nuclei in a stationary target. The outcome of the collision depends on several factors: the energy of the accelerated ion, the mass of both the projectile and the target nuclei, and the "impact parameter," which indicates whether the nuclei meet head on or strike a glancing blow. As might be expected, the most violent events are high-energy, head-on collisions between very heavy nuclei.

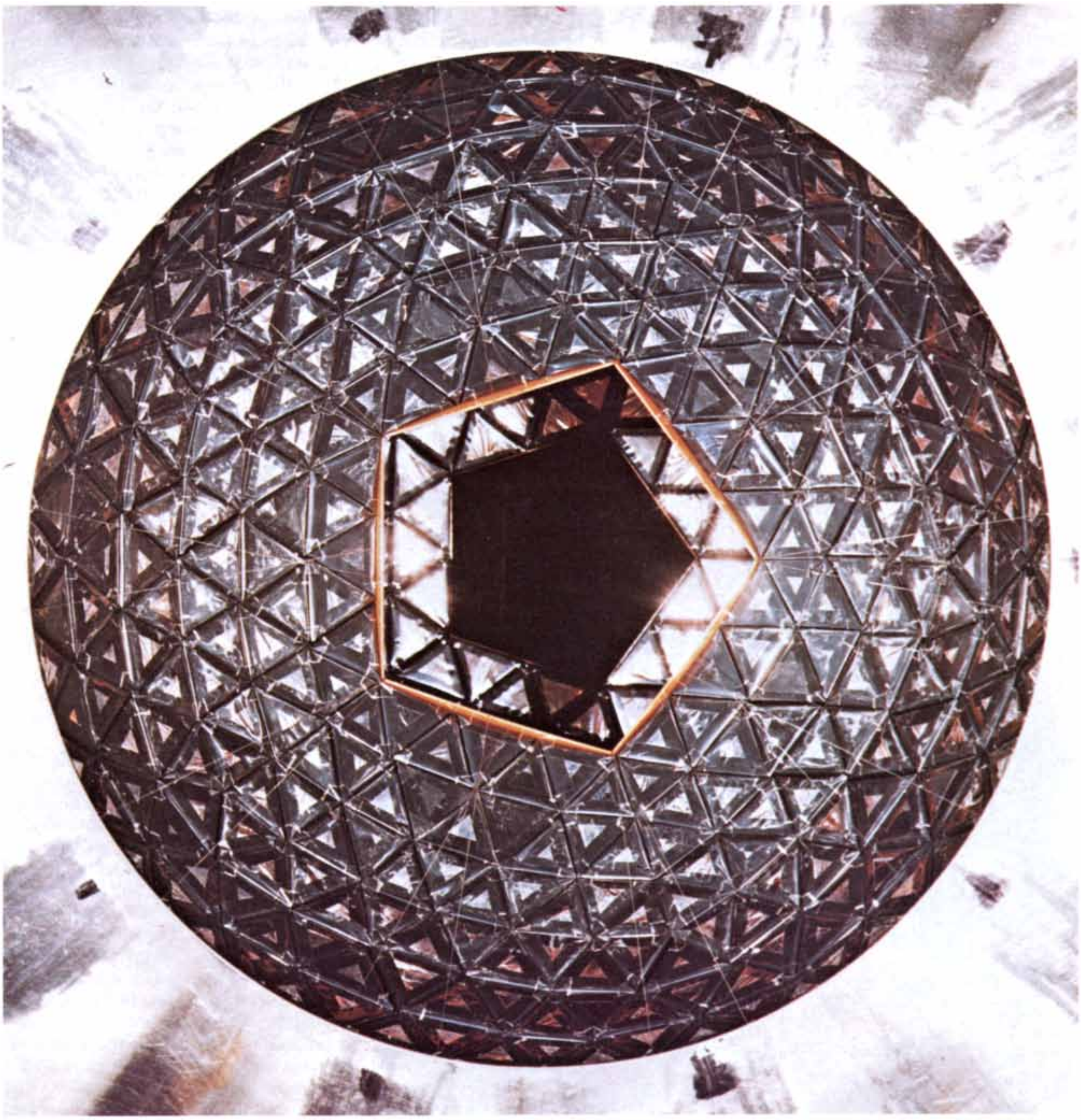
A major aim of studying such events is to determine the equation of state of nuclear matter, the equation that describes how the nucleus responds to changes in temperature, pressure and density. For an ordinary substance such as water the equation of state can be found through straightforward experiments. For example, water at atmospheric pressure goes through a familiar sequence of phases: it is a solid below zero degrees C., a liquid up to 100 degrees and then a vapor. At a much higher temperature, namely about 1,000 degrees, the water molecules are destroyed by violent collisions and water becomes a plasma made up of electrically charged ions and electrons.

Analogous phases are thought to exist in hot, dense nuclear matter. The ground-state, or normal, nucleus is rather like a liquid droplet: individual protons and neutrons (known collectively as nucleons) move about freely within the droplet but seldom stray beyond its surface. In the ground state, where the temperature is effectively zero, the density is $2.8 \times 10^{14}$ grams per cubic centimeter and the pressure is zero. (The pressure must be zero simply because the nucleus is unconfined: if the pressure were greater than zero, the nucleus would expand, and if the pressure were negative, it would contract.)

The study of heavy-ion collisions has already given evidence of a transition to a high-temperature vaporlike phase, in which the nucleons move fast enough to overcome the cohesive nuclear forces. In essence nuclear matter is observed to boil. At still higher temperatures the nucleons themselves are transformed into particles of greater mass. There is speculation that at high density but comparatively low temperature there may be an ordered phase of nuclear matter analogous to a crystalline solid. If even greater compression or heating can be achieved, the nucleons may be destroyed and the nuclear equivalent of a plasma may appear; it would consist of quarks and gluons, the constituent particles of nucleons.

Even using beams of relativistic heavy ions it is not possible to study the phase diagram of nuclear matter by means of the deliberate and carefully controlled experiments employed with other forms of matter. A high-energy collision of two nuclei is anything but controlled. The nuclei crash into each other and interact vigorously, raising the density and the temperature and thereby the pressure. There are no instrumental probes for measuring these quantities directly. Moreover, the high-density state is short-lived: in roughly $10^{-22}$ second the merged nuclei explode, scattering shards of nuclear matter in all directions. The fragments can be detected by instruments surrounding the interaction region. Whatever is to be learned from the collision must be deduced from the distribution of this debris.

In one experiment a carbon nucleus (atomic mass 12) was accelerated to the moderate energy of 70 MeV per nucleon and made to collide almost head on with a silver nucleus (atomic mass 108). The silver nucleus was in a photographic emulsion, which served as both target and detector. In the subsequent disintegration of both nuclei



**FACETED HEMISPHERE is part of an instrument called the GSI-LBL Plastic Ball, which detects the debris emitted when atomic nuclei collide at high energy. The view is of the concave, interior side of the hemisphere, looking along the path followed by a beam of accelerated heavy ions (atoms stripped of most or all of their electrons). When the detector is assembled, a similar hemisphere is joined to this one and a target is placed in the beam path at the center. The complete sphere has 815 facets, or detector elements. Each detector element consists of a thin crystal of calcium fluoride mounted on the end of a rod of transparent plastic. When an electrically charged fragment from a nuclear collision enters one of the facets, it passes through the calcium fluoride and then is stopped in the plastic, causing both materials to scintillate, or emit a flash of light. The position of the facet gives the fragment's direction; the two scintillations allow its energy and mass to be determined. Detector systems of this complexity are common in elementary-particle physics, but until recently they were rare in nuclear physics. Because they allow a collision to be reconstructed even when hundreds of fragments are emitted, they give an exceptionally clear view of the nature of nuclear matter at high temperature and pressure.**

some 16 electrically charged nuclear fragments were emitted; they were nuclei of the light elements deuterium, helium, lithium and beryllium, made up of from two to nine nucleons. They can be interpreted as particles evaporated in the boiling of the nuclear liquid [*see upper illustration on page 80*].

In another experiment the collision took place in a streamer chamber, where the passage of a charged particle leaves a bright trail of electrical discharges. An argon nucleus accelerated to an energy of 1.8 GeV per nucleon struck a stationary target nucleus of lead, again almost head on. At this much higher energy the two nuclei, which together comprised almost 250 nucleons, were completely destroyed. The fragments emitted were mainly individual nucleons and clusters of two or three nucleons. Indeed, the debris radiating from the site of the collision included more than 130 charged elementary particles, substantially more than the 100 proton charges in the original nuclei. Where did the extra units of charge come from? The collision process was violent enough to create more than 30 new particles, chiefly negative and positive pions [*see lower illustration on page 80*].

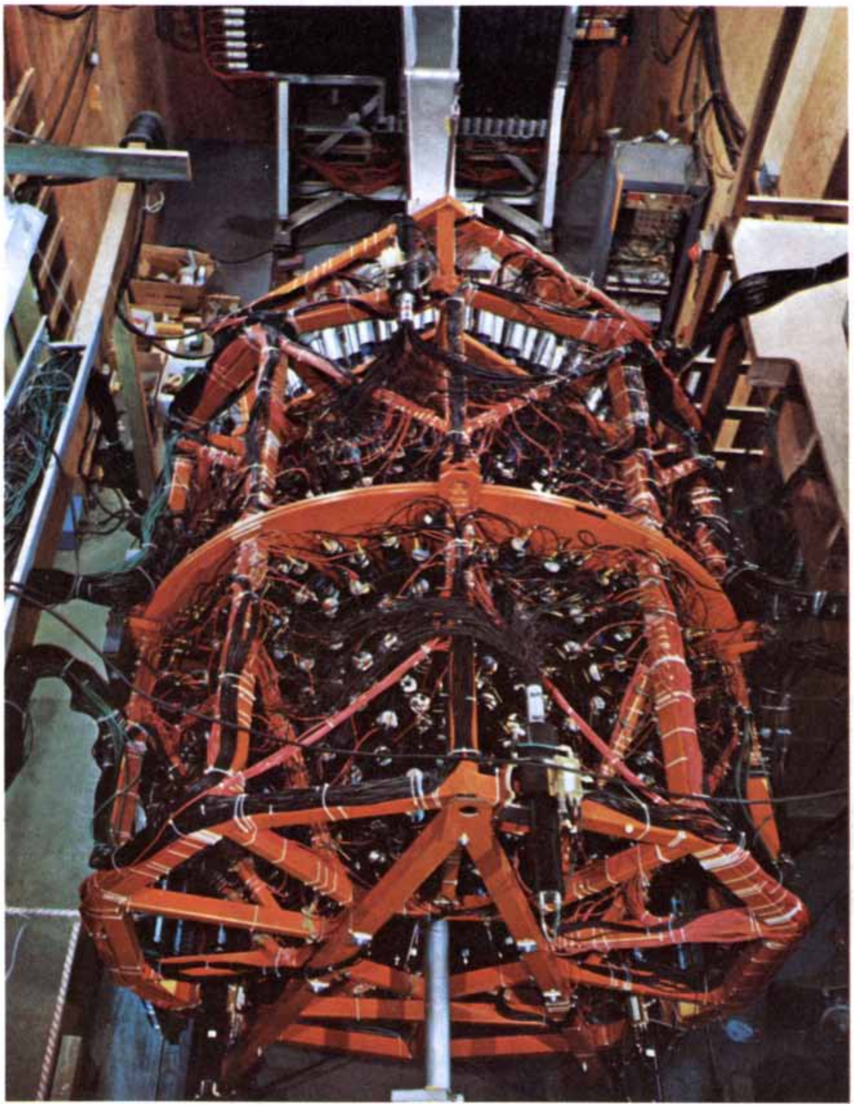
How is one to interpret such events? What are the physical processes taking place when two heavy nuclei collide at relativistic speed? The most rigorous method of analysis would be to calculate the outcome of the collision according to the principles of quantum field theory. That turns out to be impractical. Even for very small nuclei such a calculation is a mathematical tour de force; for larger collections of nucleons—in the case of uranium striking uranium there are almost 500—the task is hopeless.

Although the large number of particles rules out a quantum-mechanical calculation, it also opens the possibility of other approaches. After all, physical systems such as macroscopic solids, liquids and gases have an enormous number of component particles, and yet they are accurately described by physical theories. Indeed, a large number of particles can aid analysis by admitting models based on statistical assumptions.

Two models have taken on primary importance in the study of heavy-ion collisions. In the first model the nucleus is treated as a droplet of liquid and the equations of hydrodynamics are employed to describe the collision of two such droplets. The particles of the nucleus do not appear at all in the model, any more than individual water molecules appear in a hydrodynamic description of a raindrop. Instead nuclear matter is viewed as an ideal, continuous fluid. The hydrodynamic model was first applied to relativistic nulear collisions about 10 years ago by Werner Scheid, Hans J. Müller, Jürgen Hofmann and one of us (Greiner) of the University of Frankfort and by George F. Chapline, Michael H. Johnson, Edward Teller and Morton S. Weiss of the Lawrence Livermore National Laboratory.

In the second model, called the cascade model, the nucleons do appear explicitly; indeed, the nucleus consists of nothing else. A direct calculation of their trajectories is made possible by the simplifying assumption that each nucleon is free-moving and travels in a straight line except when it hits another nucleon. The complex forces that bind the nucleons together are ignored entirely; colliding nuclei are like colliding bags of marbles. In the cascade model the outcome of a collision is determined by simulating the paths of the nucleons with the aid of a computer; the simulation must be repeated thousands of times with different initial configurations of the nucleons. The cascade model was developed by Zeev Frankel of the Weizmann Institute of Science in Israel, Vladislav D. Toneev of the Joint Institute for Nuclear Research in Dubna, Joseph Cugnon of the State University of Liège in Belgium and their colleagues.

The predictions of the two models can be compared by following the pic-



**PLASTIC BALL is installed at the Bevalac, a heavy-ion accelerator at the Lawrence Berkeley Laboratory of the University of California. The sphere of detector elements is obscured by the photomultiplier tubes that register scintillations and by the cables that supply power to and receive data from the photomultipliers. The heavy-ion beam travels in the aluminum vacuum line that enters the detector at the bottom of the photograph. At the far end of the hall is an additional detector system called the Plastic Wall, which supplies information about the numerous high-energy fragments emitted at a small angle to the beam axis.**

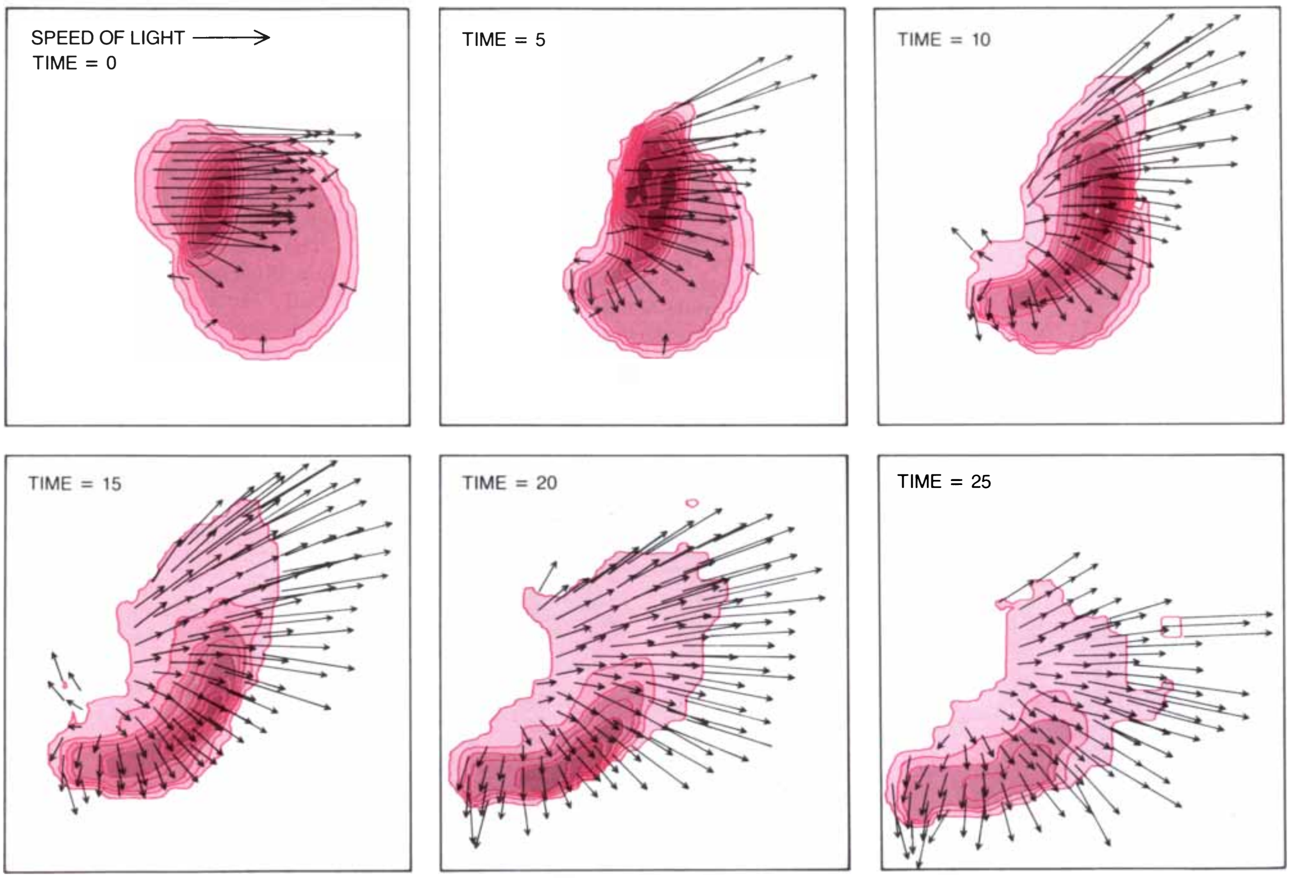
ture that each yields of a niobium-niobium collision [*see illustrations on page 82*]. In both cases the first stage is the partial interpenetration of the two nuclei, forming an ellipsoidal region of high density. As the collision proceeds more matter enters the region of compression, until all the nucleons are within a volume no larger than that of one of the original nuclei. The fused and compressed nuclei then begin to expand again, and they soon greatly exceed their original volume. Explosive disintegration follows.

The major difference between the models is observed in the later stages of the reaction. In the cascade model most of the nucleons are emitted roughly parallel to the beam axis, in the "forward" direction in the laboratory frame of reference. In spite of the high compression achieved and the thousands of individual nucleon-nucleon scatterings, the nuclei are comparatively transparent, so that many of the projectile nucleons effectively pass through the target. In the hydrodynamic model, in contrast, the nuclei stop each other. As a result of the high pressure developed there is a strong sideward peak in emission, perpendicular to the beam axis.

The nature and the origin of the sideward emission in the hydrodynamic model can be seen more clearly by observing what happens when an argon projectile strikes a target nucleus of lead [*see illustration below*]. The projectile, with an energy of 800 MeV per nucleon, moves at about 80 percent of the speed of light, which is some four times the speed of sound in nuclear matter at the ground-state density. As the projectile burrows into the target, a supersonic shock wave forms ahead of the collision region. The shock wave is analogous to the one created by a supersonic aircraft in the earth's atmosphere. At first the wave is directed predominantly forward, but as it pushes matter out of its way it begins propagating outward as well. In the late stages of the collision, when the system has reached a low density and has begun breaking up into fragments, the collective flow of matter caused by the shock wave leads to a preferential sideward emission of the reaction products.

Model calculations of the kind discussed here have been done by J. Rayford Nix and Daniel Strottman and their colleagues at the Los Alamos National Laboratory and by Gerd Buchwald, Gerhard Graebner and Joachim Maruhn of the University of Frankfort working with us.

What does nature say about the differing predictions of the two models? Do nuclei stop one another or are they transparent? First it should be noted that interpretation of the experimental results has been hampered by technological limitations. Until recently most experiments in nuclear physics were done with rather simple and inexpensive instruments. As the focus of




**SUPERSONIC SHOCK WAVE propagates through nuclear matter in a simulated collision of two atomic nuclei. The projectile, an argon nucleus made up of 40 nucleons, approaches from the left with an energy of 800 million electron volts (MeV) per nucleon; moving at about 80 percent of the speed of light, it strikes a much larger target nucleus of lead (208 nucleons). In the simulation the nuclei are represented as fluid droplets and their behavior is determined by the methods of hydrodynamics. Arrows show the velocity of small volumes of the fluid and contour lines indicate density. In the early stages of the collision the nuclear fluid is heated and compressed only in the region of interpenetration. The shock wave then develops, moving forward at first and later to the side. Ultimately, when the fused nuclei disintegrate, many fragments are emitted to the side. Times shown are in units of the time needed for light to travel one fermi, or $10^{-13}$ centimeter. The model calculations were carried out by Gerhard Graebner of the University of Frankfort.**

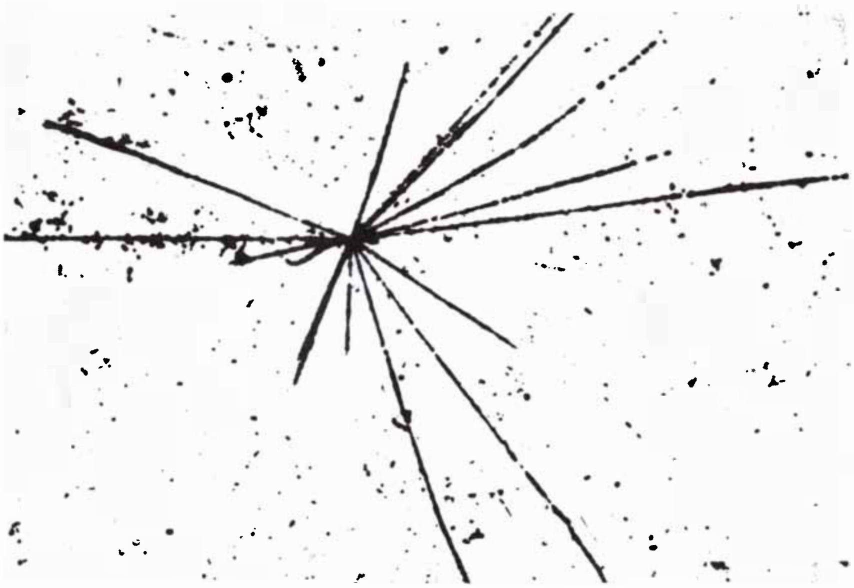
interest moves to higher energy, more elaborate detectors are needed to provide more information about the reaction products. In this case it is vital to distinguish between head-on, or central, collisions, which feature a strong supersonic shock wave, and peripheral collisions, which do not. A good criterion for making the distinction is the number of fragments emitted, but with crude detectors the selection of such high-multiplicity events can be only approximate.

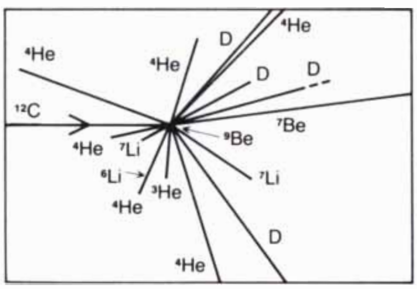
Early experiments done by Hans-Georg Baumgardt and Erwin Schopper of the University of Frankfort seemed to confirm the hydrodynamic model's prediction of sideward emission. Baumgardt and Schopper exposed solid-state detectors to beams generated by the Bevalac and by the Dubna synchrophasotron accelerator. They then selected for analysis only the high-multiplicity events and measured the angular distribution of the emitted helium fragments. They found a sideward peak above a background of isotropic, or directionally undifferentiated, emission.

To gather data of greater statistical reliability a detector based on electronic particle counters was built as a joint undertaking by a group of workers from the Gesellschaft für Schwerionenforschung (GSI) at Darmstadt in West Germany, headed by Hans H. Gutbrod and Reinhard Stock, and a group from the Lawrence Berkeley Laboratory (LBL) headed by Arthur M. Poskanzer. In high-multiplicity events (corresponding to central collisions) particles with low kinetic energy exhibited a sideward peak. The results were in agreement with hydrodynamic calculations done by Maruhn and us.

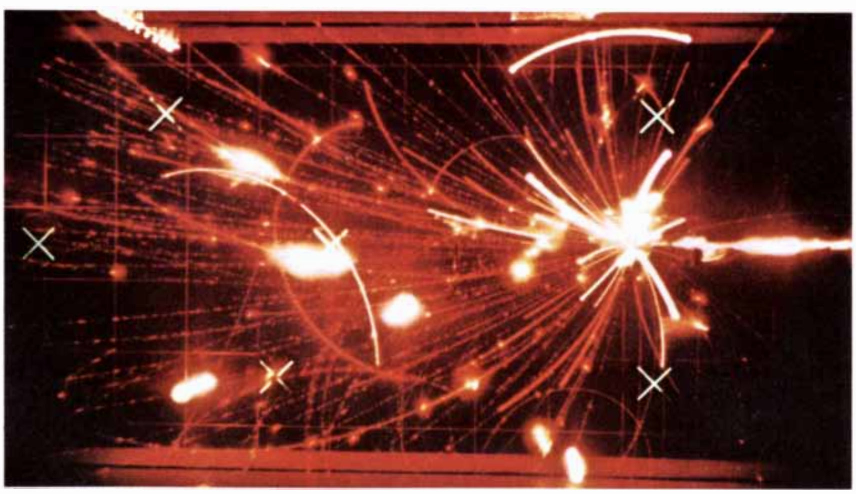
The collective sideward flow predicted by hydrodynamic calculations has been established unambiguously only in recent experiments done by Hans-Georg Ritter, Gutbrod, Poskanzer and their colleagues in the GSI-LBL group. The experiments were done with a new detector system called the plastic ball. More than 800 plastic detector elements are mounted in a spherical shell surrounding the target; they register not only the number of fragments emitted but also their mass, charge and energy. The plastic ball was first employed to study central collisions between nuclei of equal mass, where the hydrodynamic model predicts the strongest sideward emission. In collisions of calcium on calcium (atomic mass 40) no sideward peak was detected. With niobium on niobium (atomic mass 93), however, a definite pattern of sideward emission was found, in agreement with the hydrodynamic calculations made by Buchwald, Graebner, Maruhn and us.

As the situation stands now the simplest version of the cascade model seems to be limited to describing interactions of lighter nuclei. The hydrodynamic model has much broader support as a description of central collisions between more massive nuclei. Its success establishes the key mechanism for creating hot, compressed nuclear matter in the laboratory, namely the propagation of nuclear shock waves. Having such a mechanism on hand






**NUCLEAR FRAGMENTATION following a heavy-ion collision is recorded in a photographic emulsion (*above*), which acts as target and detector. A carbon nucleus accelerated to 70 MeV per nucleon has struck a silver nucleus in the emulsion. At this moderate energy some 16 fragments of nuclear matter (*labeled in the diagram at left*) radiate from the site of the collision. They are nuclei of deuterium, helium, lithium and beryllium, with from two to nine nucleons. The event was recorded by Bo Jakobsson and his colleagues from the University of Lund.**



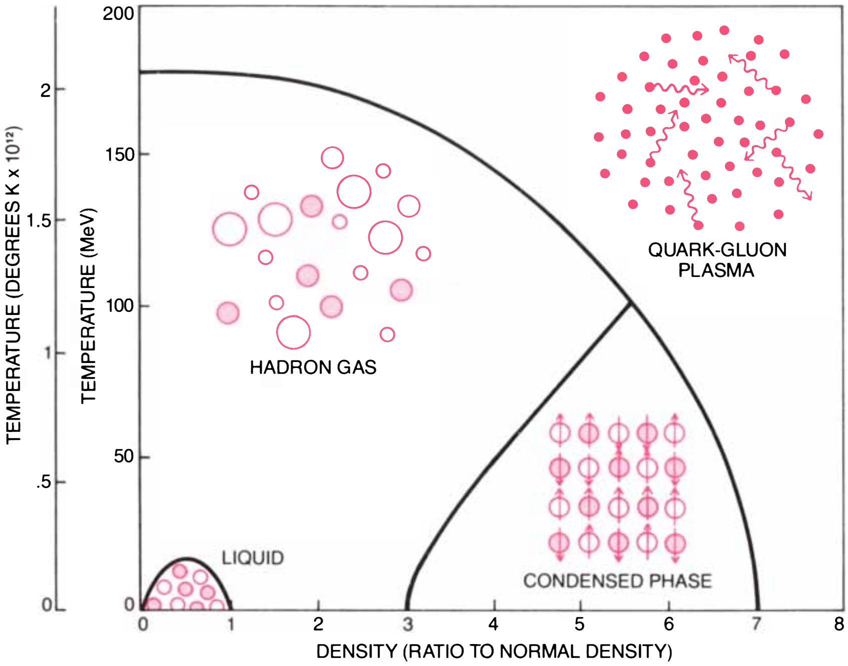
**MORE VIOLENT IMPACT of larger nuclei at higher energy leads to the complete disintegration of nuclear matter. An argon projectile at an energy of 1.8 billion electron volts (GeV) per nucleon and a lead target are broken up into individual nucleons and clusters of two or three nucleons. More than 130 fragments are emitted. In the photograph, made by Robert T. Poe and his colleagues from the University of California at Riverside, the paths of charged particles are traced by trails of electrical discharges in a streamer chamber.**

is an important stimulus to the experimental search for new, abnormal states of nuclear matter.

One indicator of a transition from the liquid to the vapor phase in nuclear matter is an increase in entropy, the measure of the disorder in a system. As the colliding nuclei are compressed the temperature rises; the nucleons are more violently agitated by their random motions. It follows that they become more disordered and the entropy increases. When the compressed matter expands again, the temperature falls, but the entropy generated in the collision cannot (according to the second law of thermodynamics) be dissipated. In other words, once the nucleon motions have become more chaotic they cannot be restored to their former order.

The production of entropy in a collision has definite effects on the debris that ultimately appears in a detector. In general as the entropy increases, the nuclear matter breaks up into a larger number of smaller fragments. This effect was seen in the two collisions discussed above. The effect of entropy on the mass spectrum has been exploited by Barbara V. Jacak, Gary D. Westfall, Konrad Gelbke, David Scott and their colleagues at Michigan State University to measure the entropy produced in collisions. At intermediate impact energies they find the entropy is considerably higher than would be expected from calculations based on the properties of normal nuclear matter.

Among various candidate explanations of the entropy excess, one possibility is particularly intriguing. The extra entropy may be a product of the phase transition from a liquid to a vapor. According to Laszlo P. Csernai of the University of Minnesota, the sequence of events could be as follows. In the early, compressive stage of the collision some of the energy of impact supplies the heat of vaporization needed to induce the phase transition. In the subsequent expansion some of the nucleons evaporate, becoming freely moving particles of a gas. As the nucleons leave the region of intense heating, however, the nuclear matter condenses into a liquid again, releasing the latent heat of vaporization. It is the liberation of this energy that gives rise to the excess entropy; indeed, the contribution to the entropy from the phase transition could be greater than that from the original heating and compression.



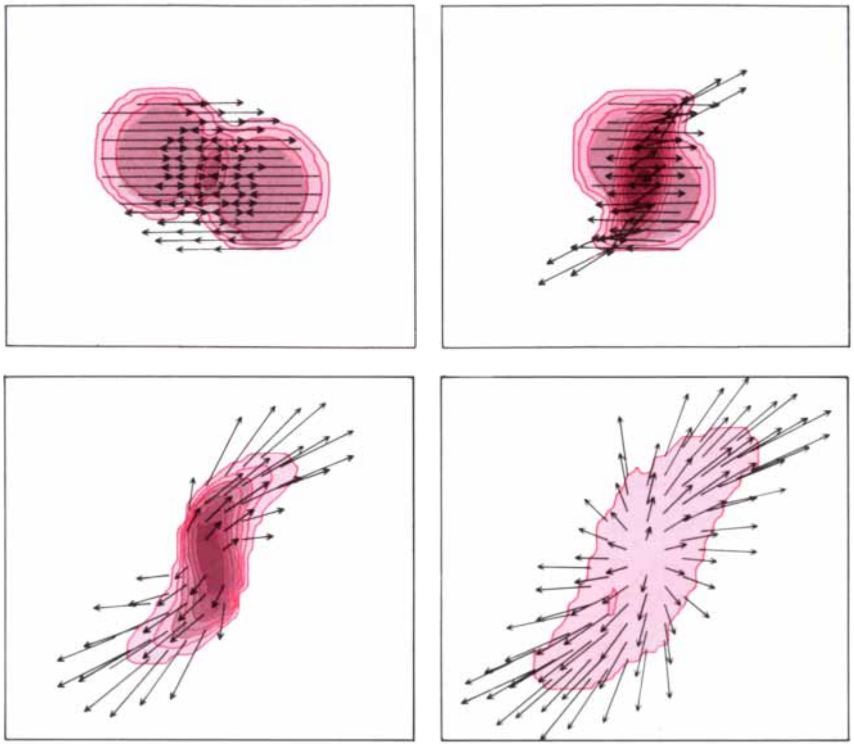

**PHASE DIAGRAM OF NUCLEAR MATTER** shows transformations predicted by various theories and conjectures. The normal, liquidlike state of nuclear matter occupies the lower left-hand corner of the diagram. At elevated temperature and density the liquid evaporates, forming a gas made up not only of nucleons but also of pions and of heavier particles known generically as hadrons. With further heating and compression the hadrons themselves could break down into their constituent particles, the quarks and gluons, creating a plasmalike mixture called quagma. At high density but comparatively low temperature nuclear matter could be frozen into a lattice structure analogous to a crystalline solid. Protons and neutrons with oppositely oriented spins would alternate in the lattice. In the pictorial representation of the phases shown here density is not indicated accurately.

The known high-density forms of nuclear matter are unstable. They can be created only by supplying energy to squeeze the nucleons closer together, and when the compressive force is released or dissipated, the nuclear matter immediately expands again. What this means in mathematical terms is that the curve described by the equation of state rises with increasing density. There is no law of nature, however, that requires the curve to continue rising indefinitely. It could have a secondary minimum, corresponding to a stable high-density configuration of nucleons. Once that configurationwasestablishedestablished,it maintained without applied pressure. Speculations about the existence of such stable "density isomers" of nuclear matter were first discussed by Arnold R. Bodmer of the Argonne National Laboratory, T. D. Lee and G. C. Wick of Columbia University and A. B. Migdal of Moscow State University.

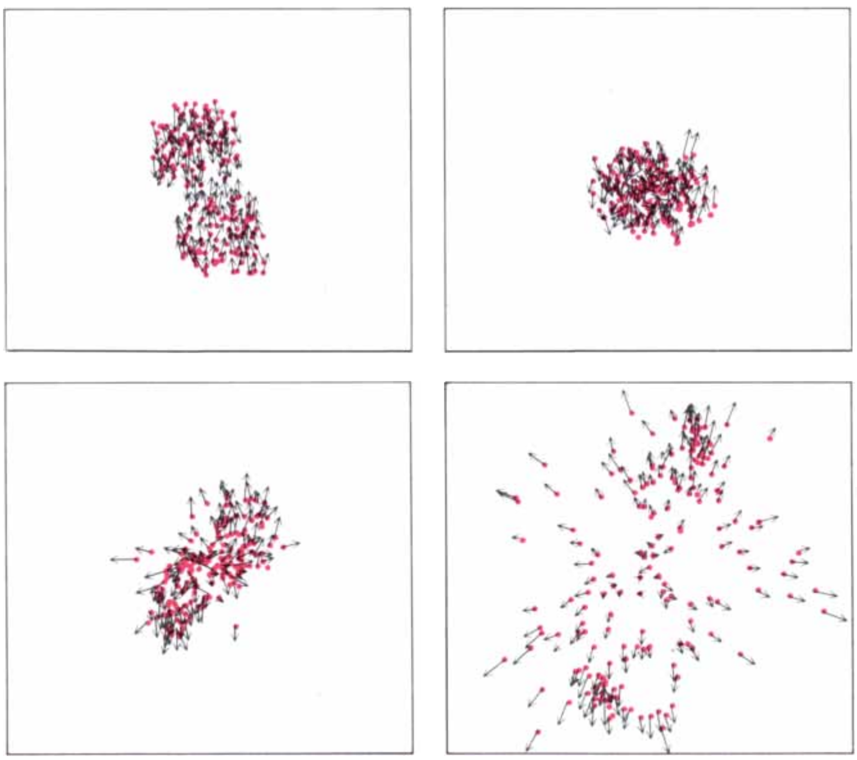
In understanding density isomers an analogy with ordinary matter is again helpful, in this instance an analogy with the solid phases of carbon. At room pressure the stablest crystalline form of carbon is graphite; it can be compressed, but when the pressure is removed, it returns to its original density. If the applied pressure exceeds a threshold, however, the carbon atoms adopt a new, denser geometric configuration, that of diamond. At low pressure diamond is metastable: it is not the lowest-energy form of carbon, but an energy barrier greatly inhibits its decay into graphite.

Lee and Wick suggested that if nuclear matter could be compressed enough to create a high-density isomer, it might form superdense "abnormal nuclei" consisting of from 300 to 100,000 nucleons. The nucleons themselves would give up nearly all their mass to a thick soup of pions (the particles that embody the nuclear force field) in which they would be embedded. Another mechanism that might lead to the formation of a density isomer was proposed by John Boguta of the University of California at Berkeley. In the strong nuclear field generated by a dense assemblage of nucleons, he pointed out, the effective mass of various excited states of the nucleon might be reduced to a value not much higher than that of the ground-state nucleon itself. Many excited states could then be produced by thermal fluctuations even at low temperature.

A third mechanism, called pion condensation, is based on calculations of the interaction between nucleons and



**HYDRODYNAMIC MODEL describes a heavy-ion collision as the fusion of two liquid droplets. Here, in calculations done by Gerd Buchwald of the University of Frankfort, two niobium nuclei collide almost head on. The collision is shown in the center-of-mass coordinate system, so that the nuclei seem to be moving with equal speed in opposite directions. The flow of matter has an important component perpendicular to the original axis of motion.**



**CASCADE MODEL of a niobium-niobium collision calculates the path of each nucleon. The model adopts the simplifying assumption that the nucleons do not interact except in two-body collisions. As in the hydrodynamic model, there is an initial stage of compression, followed by disintegration. The distribution of fragments, however, is more nearly uniform in the cascade model. Calculations are by Joseph Molitoris of Michigan State University.**

pions done by Migdal and by Gerald E. Brown and Wolfram Weise of the State University of New York at Stony Brook. Their results show that at from two to three times the normal nuclear density pions can bind nucleons into a latticelike structure comparable to a crystalline solid. Protons and neutrons would alternate in the lattice, as would the orientation of the particles' intrinsic spins.

Is there any experimental evidence of these peculiar high-density forms of nuclear matter? With Scheid we have considered (by means of hydrodynamic calculations) what events might signal the formation of the abnormal states, if they exist. We find that as the threshold for making superdense nuclei was crossed there would be an abrupt increase in the number of pions emitted. The additional pions would be given off by a nucleus as it surmounted the energy barrier and descended to the secondary minimum in the energy curve.

John W. Harris, Andres Sandoval and Stock in the GSI-LBL streamer-chamber group have recently measured the pion multiplicities resulting from collisions of argon nuclei. They have found that the pion yield increases smoothly and linearly with bombarding energy and has no abrupt jumps. They were able to extract from their data a nuclear equation of state that rises steadily with density and shows no sign of abnormal, secondary minimums.

From these findings one might conclude that the abnormal states do not exist in nature. It is also possible, however, that the accelerators of the current generation are not able to assemble a large enough collection of nucleons and maintain the required density long enough for them to cross the energy barrier, if it exists.

The several successes of the hydrodynamic model of heavy-ion collisions are in some respects surprising. The model relies on thermodynamic concepts such as temperature and pressure, which can properly be applied only to a system near thermal equilibrium. To be specific, temperature measures the vigor of the random motions of the particles in a system, but the measurement is valid only if the motions are in fact random. It is by no means obvious that this condition is satisfied in nuclear matter.

In a fluid, motion becomes random and the system reaches thermal equilibrium if the mean free path—the average distance a particle moves between collisions—is small compared with the overall dimensions of the system. Calculations of the mean free

path of a nucleon show that normal nuclear matter just barely qualifies: the mean free path is a fermi or two (one fermi is $10^{-13}$ centimeter), whereas the diameter of a large nucleus is between eight and 14 fermis. In a heavy-ion collision compression and certain subtler effects further reduce the mean free path. On the other hand, the interaction is exceedingly brief. Does the system stay together long enough for the energy of impact to be randomly distributed among the nucleons?

The justification for describing collisions of nuclei in thermodynamic terms is that the predicted temperatures are in accord with experimental results. The temperature reached in a collision can be estimated from the energy distribution of the emitted fragments. The temperatures deduced in this way agree with those calculated by hydrodynamic methods over a range of three orders of magnitude. Only at the highest impact energies does a discrepancy appear: the temperature does not increase as fast as the model suggests it should.

The flattening of the temperature curve is not a mystery; it comes about because a full description of the collision process must take into account not only nuclear physics but also elementary-particle physics. Nucleons are not inert hard spheres: at sufficiently high energy they are subject to internal transformations, much as atoms and molecules undergo chemical changes at high temperature.

The proton and the neutron are members of the large class of particles called hadrons. They are the particles thought to consist of the more elementary constituents called quarks. There are two basic groups of hadrons: the baryons, made up of three quarks, and the mesons, made up of a quark and an antiquark. The nucleons are the lowest-mass baryons and the pions are the lowest-mass mesons. It is the creation of higher-mass hadrons that limits the temperatures attained in violent nuclear collisions.

Pions are the first hadrons to appear. The rest mass of a pion is about 140 MeV, which is the threshold for pion production in proton-proton collisions. In nuclear collisions pion emission has been observed even at energies as low as 25 MeV per nucleon. Under those conditions the mechanism of pion creation clearly requires the cooperative action of many nucleons, whose total energy exceeds the threshold.

When the collision energy is greater than about 1 GeV per nucleon, hadrons heavier than pions are also created. One important process is the transformation of a nucleon into one of the particles designated $N^*$ resonances. There are many such particles, which can be arranged in a series of progressively higher mass. They all have the same basic properties as the proton or the neutron and are made up of the same constituent quarks; indeed, they can be viewed as mere excited states of the nucleons, analogous to the excited states of an atom, in which the quarks have taken on additional energy and angular momentum.

At still higher energy other mesons and baryons can be formed. Several hundred varieties of hadrons have been discovered, with masses as high as 10 GeV. In most cases the massive hadrons decay in about $10^{-23}$ second, so that there is ample time during a nuclear collision (lasting $10^{-22}$ second) for repeated cycles of creation and decay. For any given energy an equilibrium distribution of hadron types and masses is established; nuclear matter is converted into hadron matter. This is what constrains temperatures in the highest-energy experiments. Collision energy that in the absence of hadron creation would contribute to the kinetic energy of the nucleons is diverted into the mass of the excited hadrons. Instead of producing faster particles the collision forms heavier ones.

The creation of hadron matter is a kind of phase transition, and it has a latent heat, just as the liquid-vapor transition does. When liquid water is heated, the temperature rises at first, then it remains constant at 100 degrees while additional heat supplies the energy needed for vaporization; the temperature rise resumes only when all the water has been converted into vapor. The temperature curve for nuclear matter could have a similar shape. With initial heating the temperature would rise, then the rate of increase would diminish as nucleons were converted into higher-mass hadrons; the temperature would asymptotically approach some maximum value. The rapid temperature increase could resume only when all the particles in the system were in the form of the highest-mass hadron.

The trouble with this formulation is that there may not be a "highest-mass hadron." The number of hadron types seems to increase exponentially with energy. If the trend continues indefinitely, further heating will never raise the temperature of matter much beyond what has already been attained in heavy-ion collisions. There would be an absolute limiting temperature in nature. This idea was introduced by Rolf Hagedorn of CERN (the European laboratory for particle physics), who suggested the limiting temperature should

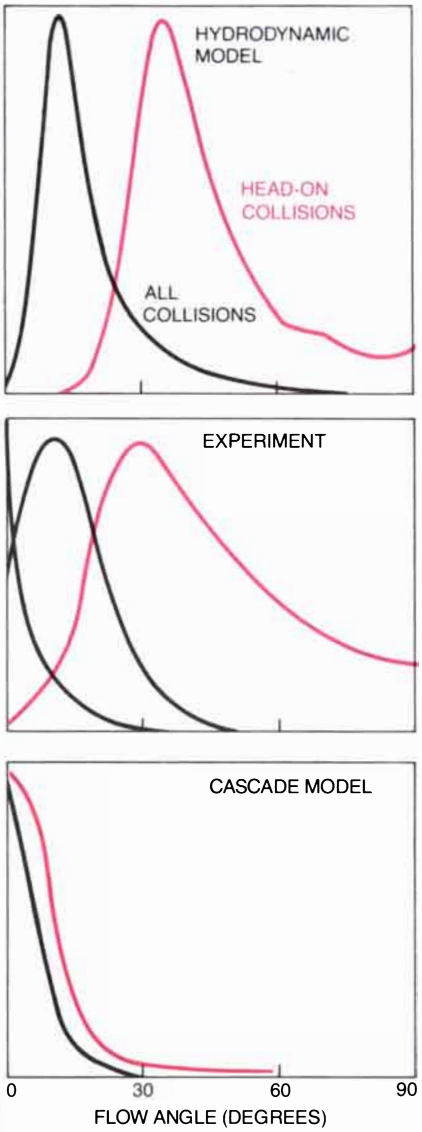
be about $1.5 \times 10^{12}$ degrees. This temperature is not far beyond what has been observed already in heavy-ion experiments.

The last of the possible phases of nuclear matter we shall consider is the one analogous to a plasma of charged particles. In it the nucleons and the higher-mass hadrons lose their identity as individual particles. The matter in such a nucleus consists of quarks and the particles that bind them together, the gluons; it is a quark-gluon plasma, or quagma.

In the laboratory the usual way to make an ordinary plasma is to heat a gas to the point where collisions between atoms are energetic enough to strip away the outermost electrons. The electrons are thereby "deconfined": they are not bound to a particular atom but can roam throughout the volume of the plasma. There is another method of achieving the same end, which operates in the interior of stars. Extreme compression of a gas can drive the atoms into close contact and even make their outer shells of electrons overlap. Under these conditions the electrons are again deconfined and can move freely from one atom to another. As in a metallic solid, a sea of deconfined electrons is shared by all the atoms.

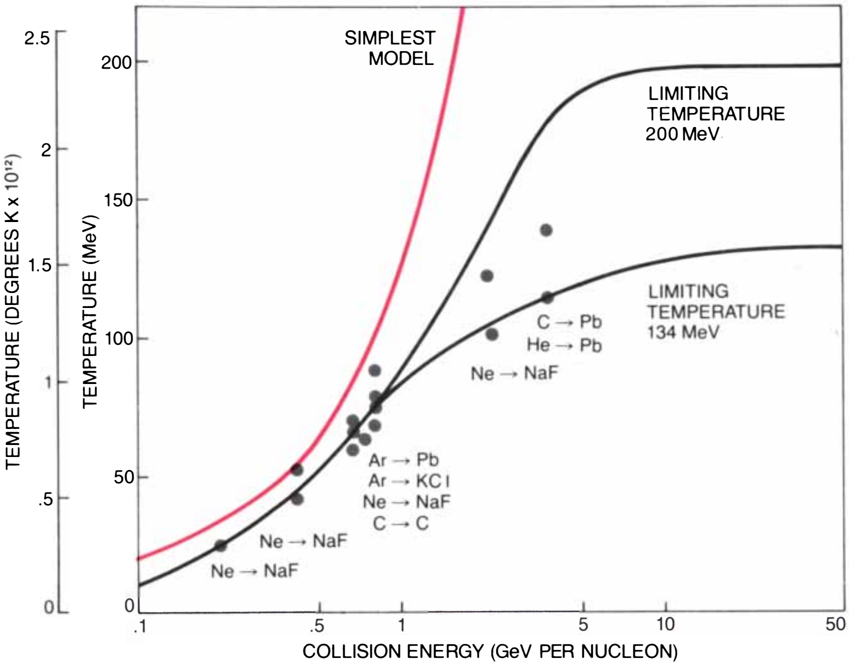
There is an important difference between the binding of electrons in an atom and the binding of quarks in a hadron. With a finite investment of energy an electron can be removed entirely from an atom. The attractive force between quarks, on the other hand, cannot be overcome no matter how much energy is supplied; the force grows stronger without limit as the quarks are separated. For this reason the quarks in a quagma cannot actually be set free; they are merely given a larger cage. In forming a quagma the aim is not to separate the quarks but to bring the hadrons closer together, as in the second method mentioned above for forming an ordinary plasma. When the hadrons are close enough to overlap, their constituent quarks can wander throughout the nuclear volume like electrons in a metal.

One obvious way to achieve the necessary density of hadrons is to compress nuclear matter until the nucleons are driven into close contact. It turns out extreme heat can also serve to create a quagma, although not by ripping quarks loose in violent collisions in the way electrons are freed in a high-temperature plasma. The mechanism is subtler: high temperature leads to the copious production of pions and other hadrons, which fill all the space between the nucleons.




**ANGULAR DISTRIBUTION of emitted fragments discriminates between the hydrodynamic (*top*) and cascade (*bottom*) models. Hydrodynamic simulations yield many fragments at a sharp angle to the beam axis; in the cascade model emission is more nearly parallel. The contrast is strongest in head-on collisions, which produce large numbers of fragments at sharp angles to the beam axis (*colored curves*). Data for the GSI-LBL Plastic Ball collaboration (*center*) support the hydrodynamic model. Head-on collisions (*colored curve*) produce more than 50 fragments at the sharpest angle; other collisions scatter 40 to 50 fragments at a pronounced angle (*middle black curve*).**

How densely must nucleons or hadrons be packed together to create a quagma? What beam energy would be needed to reach this density in heavy-ion collisions? We and others have estimated these quantities by several methods. Perhaps the simplest method proceeds from the observation that quarks seem to move freely inside a hadron. Thus if an entire blob of nu-




**HIGHEST TEMPERATURES ATTAINED in nuclear collisions signal the creation of massive hadrons and may give evidence of a maximum temperature that cannot be exceeded in nature. According to the simplest theoretical calculations (*colored curve*), temperature should increase exponentially with collision energy. Experimental data are in agreement up to about .5 GeV per nucleon, but thereafter the temperature rise is slower than expected because energy that might have heated the nuclear fluid instead goes into making hadrons of greater mass. If hadrons can have an arbitrarily large mass, it may be impossible to reach temperatures beyond some finite limit in nucleus-nucleus collisions. Temperatures based on two possible limiting values (*black curves*) were calculated by the authors.**

clear matter could be raised to the density of matter inside a hadron, the quarks should be deconfined. The critical density calculated in this way is from four to eight times the density of normal nuclear matter. This degree of compression should be readily attainable; indeed, experiments with existing accelerators, at a beam energy of 1 GeV per nucleon, probably reach densities of from three to five times the ground-state density.

Other estimates of the density needed for quagma formation are in good agreement with these. It therefore seems the required collision energy may be just beyond the reach of the accelerators operating today. A machine that could accelerate the heaviest nuclei to an energy of from 10 to 100 GeV per nucleon should be well suited to an investigation of the properties of quagma. The construction of machines in this class is under way or under consideration at several laboratories, including CERN, Dubna, GSI, the Brookhaven National Laboratory and LBL.

Some of the most interesting questions to be addressed in work with the new machines will have to do with the decay of quagma. As the density falls below the critical level the quarks must regroup to form hadrons, but at present little is known about this process, even from theoretical calculations. The mechanisms of regrouping could reveal much about the nature of the force between quarks. Perhaps a quark with high momentum can escape the quagma by pulling other quarks with it to form a hadron. Pairs of quarks and antiquarks might be created spontaneously in the wake of an escaping quark, ultimately appearing as a stream of mesons. Alternatively, the entire mass of quagma might segregate, or fracture, into quark clusters. Some of the clusters could represent unusual forms of hadronic matter.

Up to now the properties of quarks have been studied only in the interactions of individual hadrons, where the number of quarks taking part is strictly limited. Heavy-ion collisions offer the possibility of observing many-body quark interactions. By bringing together several hundred quarks and gluons, quagma can be studied not as a collection of independent particles but as a form of matter. It is the form of matter with which the universe began, in the first instants after the big bang, and so the highest-energy heavy-ion reactions could well reenact at small scale the early evolution of the universe.

# The Reliability of Computer Memories

*Large ones consisting of hundreds of silicon chips are inherently liable to fail unless steps are taken. The steps are mathematical, based not on preventing errors but on correcting them afterward*

by Robert J. McEliece

All modern computers, from the most modest microcomputer to the grandest supercomputer, have electronic memories that are built from silicon chips. The technology is highly reliable: any one memory chip can be expected to function faultlessly for decades. But when hundreds or even thousands of chips are combined in the memory of a large computer, the time until one of them fails can be as short as a few hours. For a computer serving business or science that is plainly unacceptable.

The solution to the problem is derived from mathematics and is called an error-correcting code. It enables a memory to function perfectly even after hundreds of errors; indeed, the average time until the memory suffers so many errors that the code can no longer cope is typically many years. The code turns a highly unreliable memory into a highly reliable one; there appears to be no way to build reliable large memories without it.

How do errors occur? A memory chip is a square array of data-storage cells. The most popular variety is the 64K chip, where K, short for kilo-, means not 1,000 but $2^{10}$, or 1,024, the closest power of 2. The 64K chip stores 65,536 bits, or binary digits, of data. It costs about five dollars. The 256K chip, which stores four times as much, is even now replacing it in popularity. A one-megabit chip is rumored. (A megabit is $2^{20}$, or 1,048,576, bits.) In the 64K chip the array of data-storage cells is 256 by 256. Each cell stores one bit: a 0 or a 1, which together constitute the two-letter alphabet by which information is coded in computers.

The chip is random-access; that is, the content of each cell can be stored, or "written," and retrieved, or "read," individually. For that purpose each cell must have a unique address. The address has two parts: the row address, a binary number that identifies the cell's horizontal placement in the array, and the column address, a binary number that identifies the cell's vertical placement. The rows and columns are numbered beginning with 0; accordingly the largest address in a 64K chip identifies the cell in the 255th row and the 255th column. Now, the binary equivalent of 255 is 11111111, a sequence of eight bits; thus the row and column addresses for a 64K chip require eight bits each, or 16 in all. A 256K chip has 512 rows and 512 columns and so requires 18 bits to address a cell. A one-megabit chip (when it arrives) will have 1,024 rows and 1,024 columns and will need 20-bit addresses.

The 0's and 1's stored in a memory chip are represented by the presence or absence of negative electric charge at sites in the silicon crystal whose electrical properties make them potential wells, or electronic traps for negative charge. When a 0 is to be stored in a given cell, the potential well at the site is filled with electrons; when a 1 is to be stored, the well is emptied of electrons. Then when the cell is read, its negative charge is measured. If the charge exceeds a certain value, the stored bit is declared to be a 0; otherwise it is declared to be a 1.

If a potential well were to somehow lose its charge, or if an uncharged potential well were to somehow acquire a charge, the stored bit would be erroneous. Such mishaps do occur. In what are called hard errors the chip itself is damaged. A given cell can become unreliable, stuck at 0 or stuck at 1 or switching erratically between 0 and 1. A defect in the addressing hardware can make the data in a row or a column unreliable. Half chips and even entire chips have been known to go bad. In what are called soft errors the chip itself suffers no permanent damage. Here the usual villain is the alpha particle: the helium nucleus (consisting of two protons and two neutrons) ejected from heavy atomic nuclei during radioactive decay.

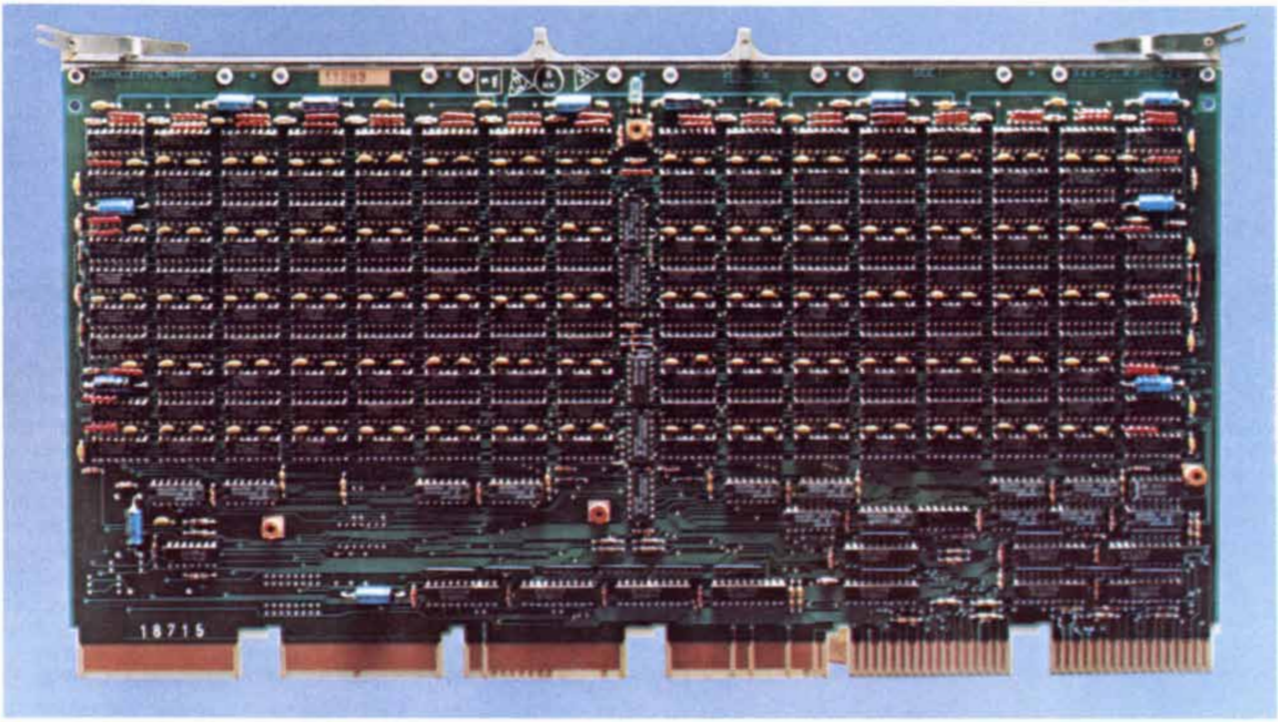
Alpha particles are distressingly common in computer memories because radioactive nuclei are found in small quantity in nearly all materials. In particular, uranium and thorium atoms are found in the plastic packaging of a typical memory chip. Thus the chip itself is subjected to constant bombardment. When an energetic alpha particle penetrates a potential well, it dislodges electrons from their position in the silicon crystal lattice. The electrons can be attracted to nearby potential wells. If a given well is already filled with electrons (that is, if a 0 is stored), this causes no problem. If, on the other hand, the well is initially empty (that is, if a 1 is stored), an error can result: the 1 can be changed to an erroneous 0.

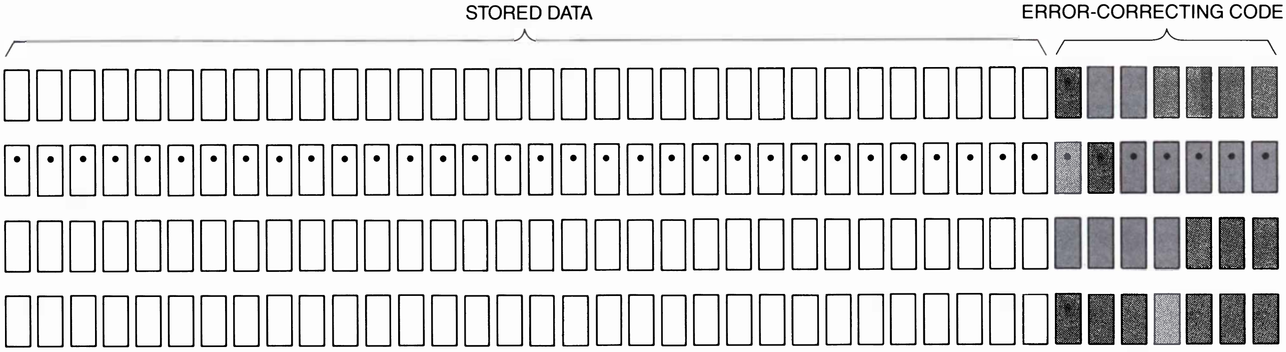
How does an error-correcting code cope with this situation? The following analysis of an imaginary "one megabyte" memory consisting of 128 64K chips provides the answer. In such a memory the chips are arranged in four rows of 32 chips each. Since each chip contains 65,536 memory cells, the one-megabyte memory has a total of 8,388,608 cells. (By today's standards for computer memories that number is quite small.) Data are divided into "words" of 32 bits each. Each word consists of the content of one memory cell in each of the 32 chips in one of the rows.

How long can this memory function before it fails? Well before the computer age, statisticians introduced the notion of mean time before failure, which signifies the average time until a device that was working perfectly fails. The mean time before failure for a memory cell in a 64K chip is gratifyingly long. Laboratory experiments (ones that take into account soft errors only) suggest that a typical value exceeds a million years. Still, as I have noted, a computer memory can have millions of memory cells, which implies that the net mean time before failure is dangerously low. For the sake of argument let the mean time before failure for a single memory cell be exactly one million years. Then the mean time before failure for a one-megabyte memory is one million years divided by 8,388,608, or about 43 days. It is far too short a time, and yet there seems to be no economically feasible way to shield a computer memory from alpha particles.

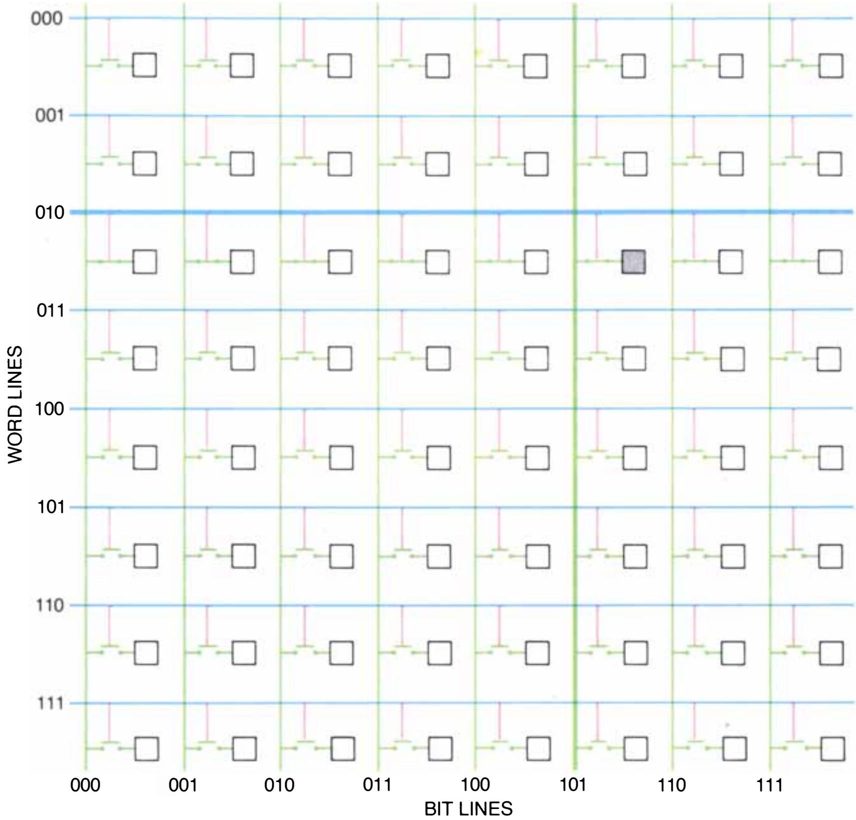
The solution to the dilemma comes from mathematicians, who, motivated by the needs of sophisticated telecommunications systems, discovered more than 30 years ago that in some circumstances it is better to correct errors than to try to prevent them. A small-scale example will explain this strategy. Suppose you want to store four bits of data—say 1110—in a memory you know to be error-prone. To represent the memory I shall employ a Venn diagram consisting of three mutually intersecting circles [*see illustration on page 92*]. Such diagrams are employed in mathematics to represent relations among sets. The three intersecting circles form a total of seven compartments.

Assign the bits to the innermost four compartments of the diagram. Then fill the remaining three compart-

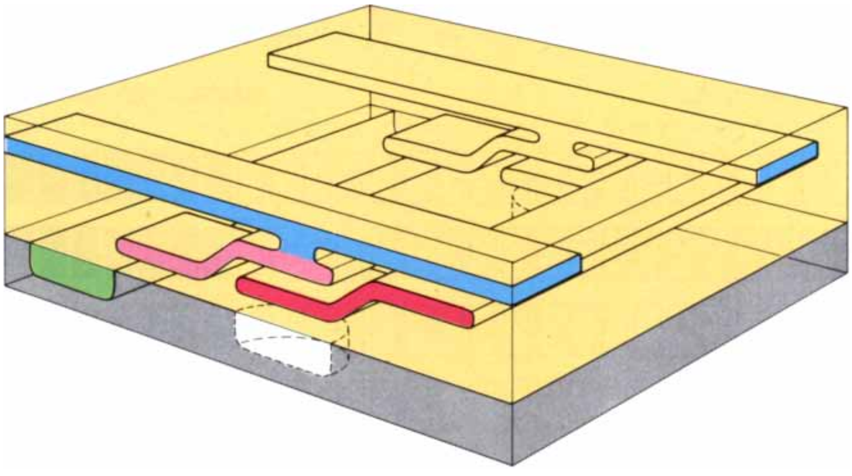





**ONE-MEGABYTE MEMORY stores $2^{23}$, or 8,388,608, binary digits of data in silicon memory chips, which fill the upper two-thirds of the memory circuit board (*top photograph*). Each chip has a capacity of 64K; that is, it stores $2^{16}$, or 65,536, bits. Physically the chips are arranged in 10 rows and 16 columns; four locations are occupied not by a chip but by a condenser (*blue cylinder*). Electronically the chips are in four rows of 39 (*bottom drawing*). The memory organizes data in units of words; each word, consisting of 32 bits, requires one location, or memory cell (*black dots*), in each of 32 chips. The remaining four rows of seven chips (*gray*) are reserved for the storage of bits for error correction. The chips at the bottom of the circuit board serve auxiliary functions such as communication and timing.**




**MEMORY CHIP IS ORGANIZED** as an array of memory cells, each cell accessible to the computer so that individual bits can be stored or retrieved. Here the organization is diagrammed as an array of eight cells by eight; in a 64K chip the array is 256 by 256. The array is square, a strategy that minimizes the required amount of wiring. To access a cell, signals are directed through the two wires intersecting at the cell. In the drawing the row address 010 is employed to activate a word line (*blue*); this closes the gates in the row. Then the column address 101 is employed to activate a bit line (*green*). The bit line "reads" the content of a cell (*gray*) by sensing electric charge in the cell. It can also "write" a bit.



**INDIVIDUAL MEMORY CELL** in a silicon memory chip is shown in schematic cross section. An electrode (*dark red*) with a permanent positive charge creates a "potential well" (*white*)—in effect a trap for electrons—in the underlying silicon crystal lattice (*gray*). The silicon is a *p*-type semiconductor, which ordinarily does not harbor free electrons. The word line accessing the cell (*blue*) is a metal strip insulated by oxide (*brown*). The bit line accessing the cell is *n*-type silicon (*green*), a semiconductor that does harbor free electrons.

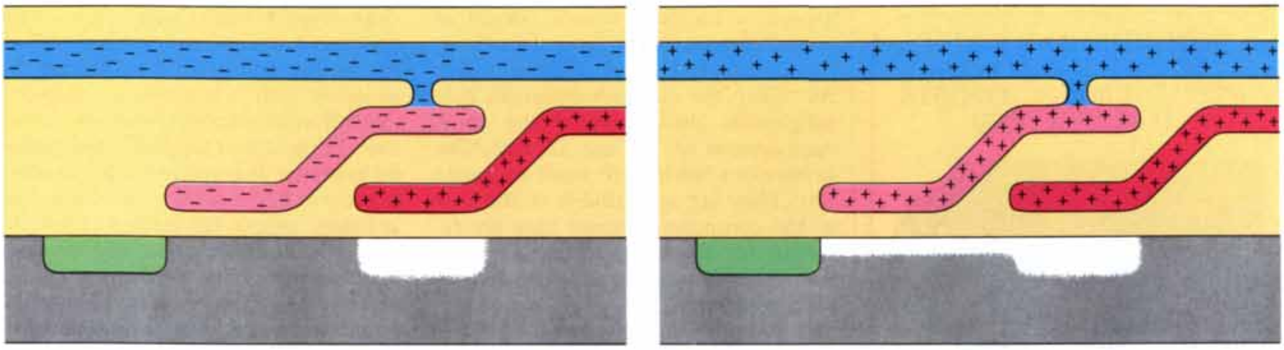
ments with what are called parity bits. The rule for generating the parity bits is that the total number of 1's in each circle must be even. For example, circle *A*, at the upper left of the Venn diagram, includes as its data three 1's. Accordingly the parity bit in circle *A* must be a 1. When the procedure is complete, each of the seven compartments contains one bit. The seven bits are called a codeword; the procedure producing the codeword from the four-bit dataword is called an encoding algorithm.

If one of the bits in the codeword now suffers an error—if a wayward alpha particle changes a bit from a 1 to a 0—the error is readily found. One need only check the parity bits. In the instance shown in the illustration the bits reveal that something is wrong in circle *A* and circle *C* but not in circle *B*. One and only one of the seven compartments has the property of being in *A* and *C* but not *B*. Hence the bit in that compartment is changed. In this way the error is corrected. The procedure is called a decoding algorithm.

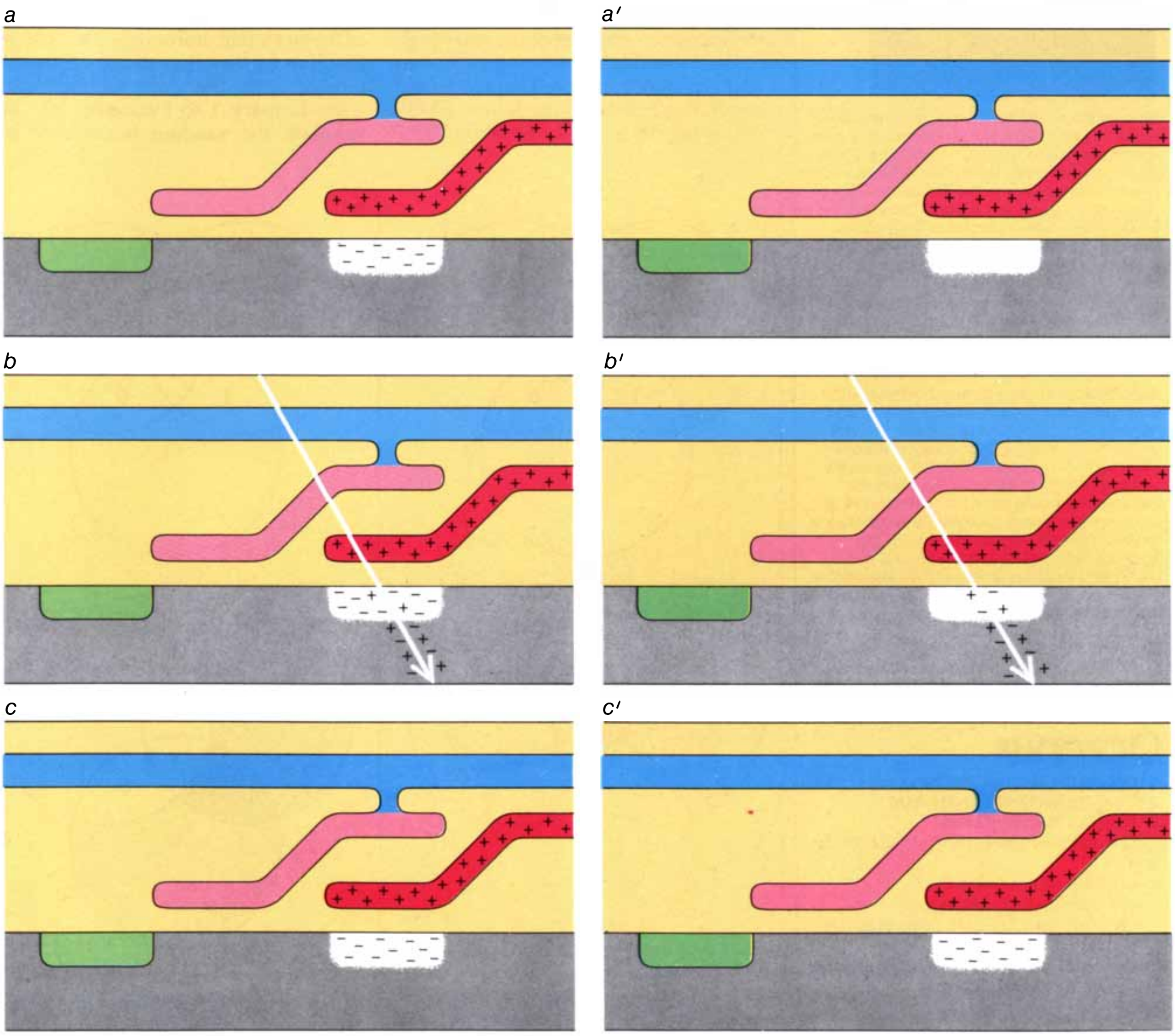
To be sure, the procedure has a limitation: it succeeds only when the codeword has suffered one error. If two or more errors occur, the procedure cannot succeed. In fact, if precisely two errors occur, the decoding algorithm will misdiagnose the problem as a single error and its effort to remedy things will make matters worse by introducing a third error. This undesirable activity can be remedied, if only partially, by means of a further parity bit outside the intersecting circles [*see top illustration on page 94*]. The further bit is chosen so that the total number of bits throughout the diagram is even. It does not enable the decoding algorithm to correct dual errors, but it does enable the algorithm to discover that something is wrong and do no harm.

The technique I have been describing was discovered in 1948 by Richard Hamming at the Bell Telephone Laboratories, and in his honor it is known as Hamming code. In particular, the scheme in which seven-bit codewords each protect four bits of data is the (7,4) Hamming code; the scheme that requires a further bit is the (8,4) Hamming code. There are many other varieties: the (15,11) code, the (16,11), the (32,26), the (64,57), the (128,120) and so on. Although they cannot easily be depicted with Venn diagrams, they are readily implemented in high-speed digital hardware.

I can now return to the question of the reliability of a one-megabyte memory. The use of a Hamming code to correct errors means that a real one-



**READING OR WRITING OF DATA** in an individual memory cell occurs when the word line closes the gate to the cell, establishing communication between the cell and its bit line. In the drawing at the left the word line is negatively charged. As a result the gate is open. In the drawing at the right the word line is positively charged. As a result the gate is closed: the charge extends the potential well, allowing communication between the cell and the bit line. The charge in the cell can hence be sensed or adjusted.

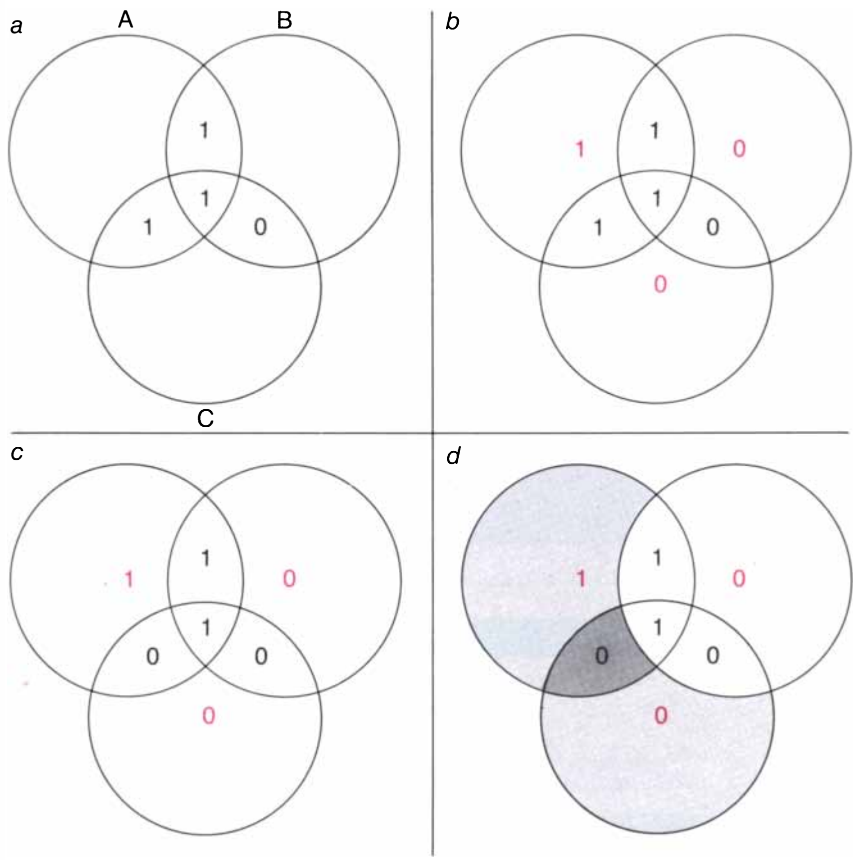



**"SOFT" ERROR** affects the reliability of a stored bit without damaging the memory chip itself. Here one cause of soft errors is shown. A bit has been stored as charge or the absence of charge: a 0 is represented by electrons in a cell (*a*); a 1 is represented by the absence of electrons (*a′*). Then the error occurs. A radioactive impurity in the plastic housing of the chip emits an alpha particle (a helium nucleus), whose passage through the chip dislodges electrons from their position in the silicon crystal lattice (*b,b′*). In the case of a 0 this causes no problem (*c*). In the case of a 1 the electrons attracted toward the electrode positioned over the cell produce an erroneous 0 (*c′*). "Hard" errors also occur; for example, damage to addressing hardware can affect entire regions of a chip.

megabyte memory cannot consist of only 128 64K chips. In a likelier arrangement 156 chips are employed. As before, the memory organizes data into words. Now, however, the words each consist of 39, not 32, bits. The seven extra bits in each word are parity bits. They are unavailable to the user of the computer. Instead they are reserved for the error-correcting code. The scheme employed is the (64,57) Hamming code. That is, the seven parity bits could actually protect 57 bits in codewords of 64 bits. The architects of modern computers have settled, however, on a standard of 32-bit words. Thus the parity bits protect fewer data bits than they could.

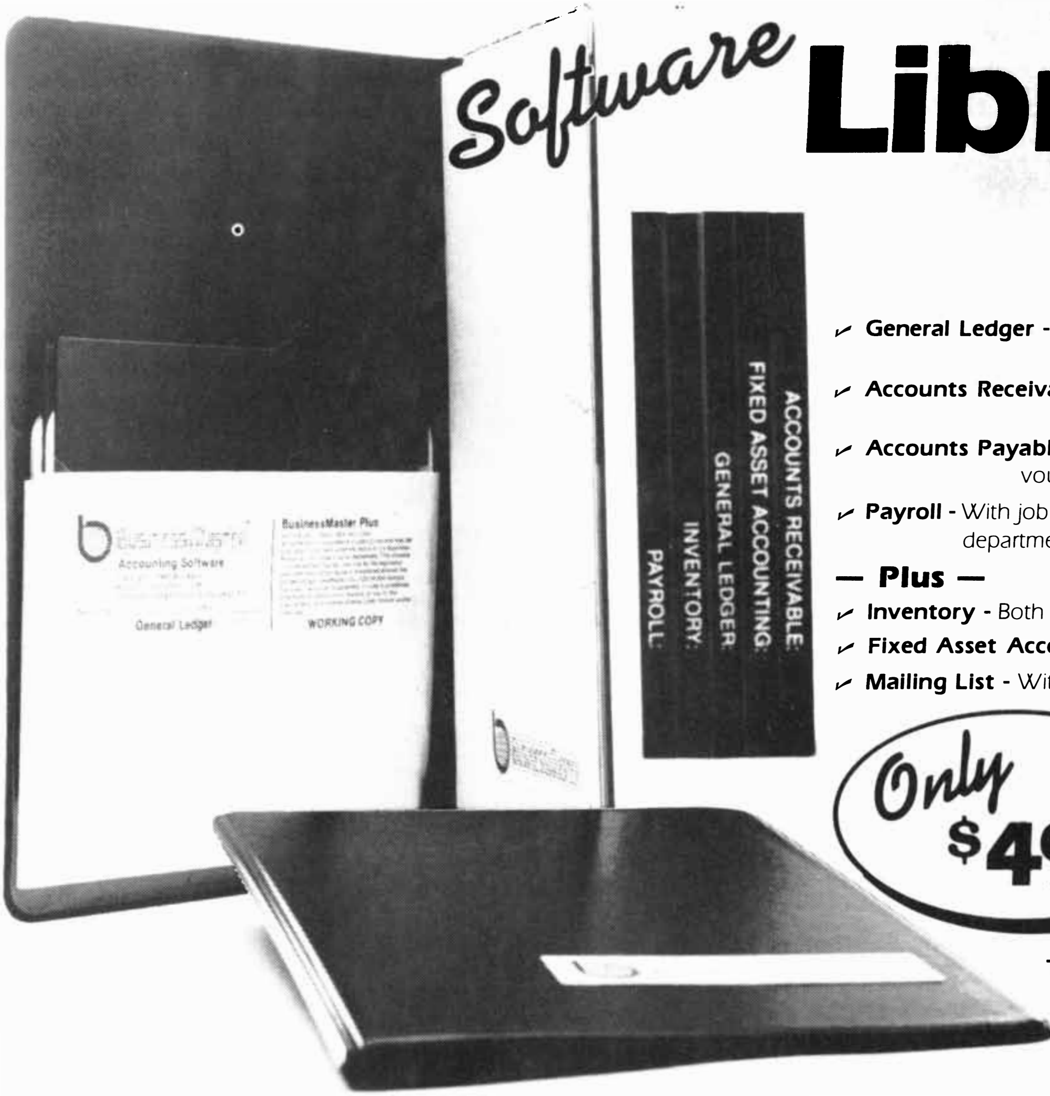
How much more reliable is a memory with an error-correcting code than a memory without it? On the one hand, the increased number of memory cells to accommodate parity bits means that errors occur more frequently: the mean time before failure falls to 32/39 of its previous value, or about 35.7 days. On the other hand, each 39-bit codeword can now tolerate a single error in any one of its bits. Hence the memory fails when some codeword has suffered two errors—but not before that. In particular the first error causes no problem; it is correctable wherever it occurs. The second error causes no problem unless the error happens to strike the codeword that already contains an error. The third error causes no problem unless it strikes either the first or the second damaged codeword, and so on. Since the memory includes 262,144 codewords, a long time is likely to pass before some codeword is hit twice.

How long? The question was answered some time ago, in a totally different context. Imagine a computer memory that stores only 365 words. Imagine further that these words are labeled with the days of the year, from January 1 to December 31. To simulate the random occurrence of




**(7,4) HAMMING CODE is the simplest of the error-correcting procedures devised by Richard Hamming at the Bell Telephone Laboratories and employed today to protect data in computer memories. The (7,4) code requires three so-called parity bits to protect each four bits of data. The construction called a Venn diagram depicts the scheme. Four bits of information are stored in the four central compartments of the diagram (*a*). Then three parity bits (*color*) are stored (*b*). The rule is that the total number of 1's in each circle must be even. A soft error changes one of the data bits (*c*). The error is detected by reexamining the parity bits, which reveal (*d*) something wrong in circle *A* and circle *C* but not in circle *B*.**

# The Complete Accounting *Software* Library




- **General Ledger** - With check disbursements & a full client write-up
- **Accounts Receivable** - With order entry & invoicing
- **Accounts Payable** - With purchase orders, vouchering & cash requirements
- **Payroll** - With job costing, multiple deductions, departmentalization

**— Plus —**

- **Inventory** - Both finished & raw goods
- **Fixed Asset Accounting** - With Cladr & ACRS
- **Mailing List** - With label printing capabilities

*Only* **$495**

**Technical Support, with Purchase**

Compiled Under: CB-80/86 Using Digital Research Access/Display Manager
Runs Under C/PM 2.2, C/PM 86, M/PM86, MS/PC-DOS
Available In Formats To Fit Most 8 Or 16 Bit Machines

*New*

## Minimaster — $159

Does everything "Stars Accounting Partner" or Rocky Mountain's "Desk Top Accountant" packages do, for less

## Maximaster — $289

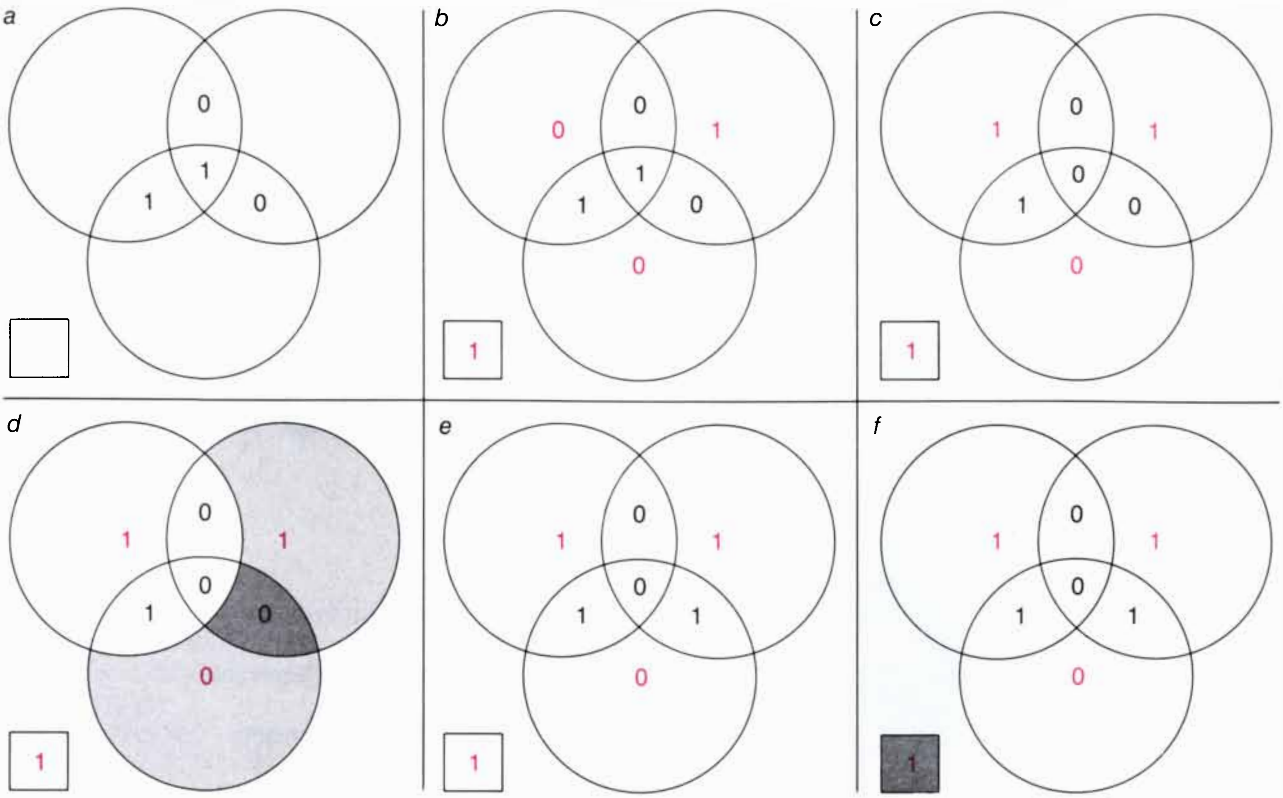
Does everything "Peach Tree Accounting" does and more, for less

*order now* **(619) 434-6165**

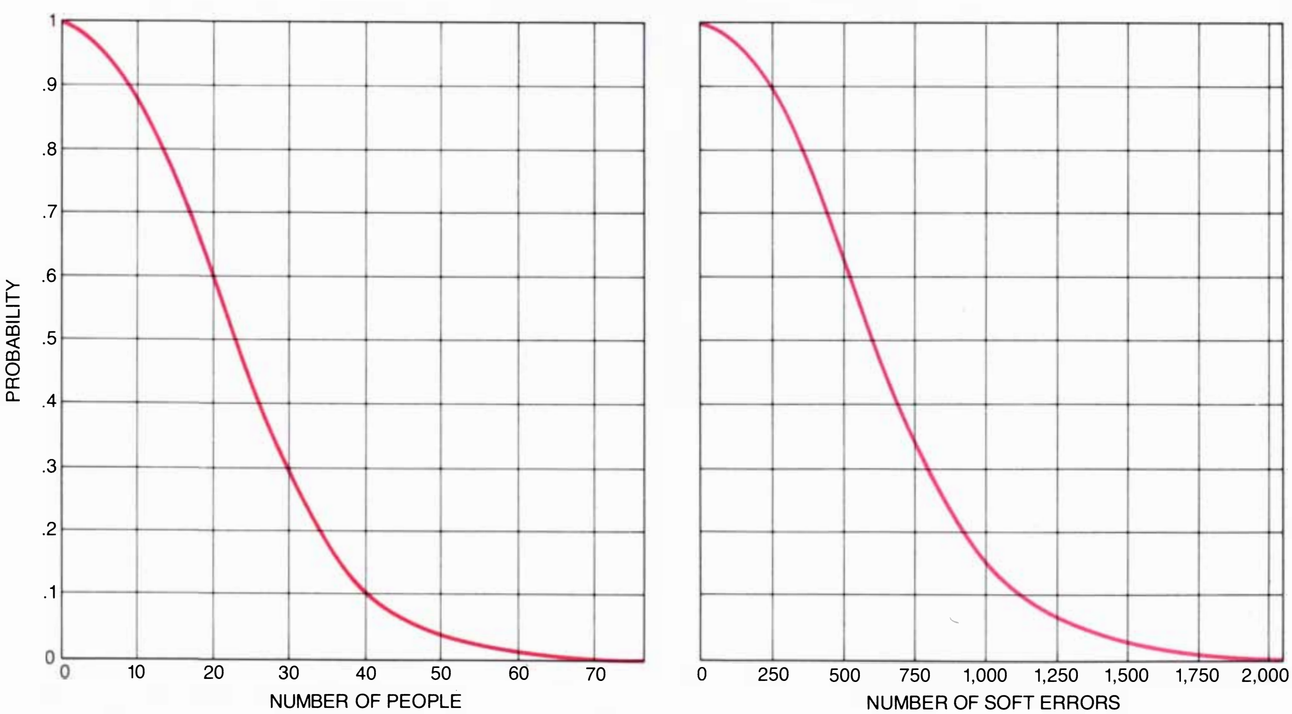
**P.O. Box 2326 Carlsbad, Calif. 92008**

BUSINESSMASTER®

BusinessMaster Registered Trademark Of BusinessMaster
*Access/Display Manager Registered Trademark Of DRI




**(8,4) HAMMING CODE** employs a further parity bit and enables the computer to detect the occurrence of dual errors, if not to correct them. Again four data bits are stored in the central compartments of a Venn diagram (*a*). Three parity bits (*color*) are placed in the outer compartments (*b*). A further parity bit is added (*box*). The rule is that the total number of 1's in the diagram must be even. Two soft errors occur (*c*). Under this circumstance the checking procedure goes astray (*d*) and worsens the problem by creating a third error (*e*). The extra parity bit is now in error (*f*). The entire sequence of bits can therefore be tagged as unreliable.




**BIRTHDAY PARADOX** can be used to predict the reliability of a computer memory protected by a Hamming code. The chart at the left shows the probability that a given number of people will all have different birthdays. In a group of 40, for example, the probability of no birthday duplication is only about .11. The chart at the right shows the probability that a 256K computer memory will function reliably in spite of single-cell errors. Since single-cell errors are counteracted by a Hamming code, the memory fails only when two errors occur in any one "word." Both of the probability curves are well approximated by the same mathematical formula.

errors in this memory you perform the following experiment. You interview a number of people and score errors according to their birthday. If, for instance, you encounter someone whose birthday is April 28, you tally an error for the computer word labeled April 28. The memory fails as soon as some birthday shows up twice. Probability theorists have shown that on the average you need interview only 24.6 people before finding two with the same birthday. This rather surprising fact is called the birthday-surprise paradox. It follows that on the average the 365-word memory will tolerate 24.6 errors before failing.

That figure depends, of course, on the total of 365 possible birthdays. Mathematicians long ago calculated the corresponding figures for an arbitrary number of possible birthdays. The general rule is expressed by a rather complicated equation. Still, on a planet with $N$ days in its year the average number of creatures you will have to interview until you find two with the same birthday is closely approximated by the square root of ($\pi/2 \times N$). The determination of the mean time before failure for a one-megabyte memory protected by error-correcting code can now be completed. The memory consists not of 365 but of 262,144 words; as a result the memory will tolerate an average of 642 errors. If an error occurs every 35.7 days, on the average, the mean time before failure is 642 × 835.7, or about 63 years.

In brief, without error-correcting code the mean time before failure is a few tens of days; with the code it is a few tens of years. The contrast illustrates the remarkable power of error-correction. To be sure, in analyzing the reliability of computer memories I have considered only "soft," single-cell errors. "Hard" errors also play an important role in determining the mean time before failure. Hard errors too can be analyzed by means of "birthday theory." They in no way alter the conclusion that error-correcting code is essential for the construction of reliable computer memories based on silicon chips.

Thus coding theory, a mathematical discipline devised in the late 1940's in response to the needs of the telecommunications industry, finds an unexpected application in the manufacture of present-day computers. Moreover, the analysis of the advantage of error-correcting code for computer memories benefits from birthday theory, an aspect of pure—that is, "impractical"—mathematics. I find these examples quite typical of the value of basic research.

# Jet-propelled Swimming in Squids

*A swimming squid takes up and expels water by contracting radial and circular muscles in its boneless mantle wall. Elastic collagen "springs" in the muscle increase the power of the jet*

by John M. Gosline and M. Edwin DeMont

The animals that run, swim or fly the fastest and the farthest are vertebrates and insects. Their locomotive prowess is attributable to rigid skeletons, built of stiff materials such as bone or cuticle, which support muscle contractions that are then translated into ordered movements with the help of flexible joints. In contrast, soft-bodied animals, whose skeletal support derives from hydrostatic pressure, usually lack the muscle power for sustained rapid movement; many of them are sedentary.

Squids are a remarkable exception. They swim as well as many fish, but they do so by jet propulsion: they suck water into their cylindrical mantle cavity through openings on each side of the head and then expel it under high pressure through a funnel. This swimming technique can be adapted for both speed and endurance. The Japanese squid, *Todarodes pacificus,* migrates 2,000 kilometers, swimming continously for two and a half months at an average speed of .3 meters, or .9 body lengths, per second. Over short distances squids have been clocked at three meters per second, and some large species are thought to swim three times as fast. Very small squids such as the larvae of *Loligo vulgaris* are capable of explosive bursts in which they move 25 body lengths in a second. With a rapid sequence of contractions squids can accelerate away from a predator, leaving behind only a cloud of ink.

They can also descend swiftly on prey. Squids are active hunters, and their way of life is made possible by physiological features resembling those of fish more than those of other mollusks such as clams or snails. They have a complex central nervous system and eyes that are nearly as developed as those of human beings. A high-pressure, closed circulation and a well-developed respiratory system enable them to sustain the high metabolic rate they need in order to jet. Thus on one level it is not surprising that squids compete well with fish.

Yet it remains interesting that they are able to do so without the benefit of a vertebrate skeleton. The bones in a vertebrate skeleton are rigid structures that act as levers to translate muscle contractions into organized movements. A hydrostatic skeleton such as that of a squid, in contrast, is not intrinsically stiff; it is essentially a tissue container filled with a working fluid. Because the fluid is incompressible, its volume cannot change when muscles in the body wall (the container) contract. If the muscles contract in balance, the fluid is pressurized and the skeleton becomes rigid, but if one muscle group dominates the others, the animal changes shape and moves.

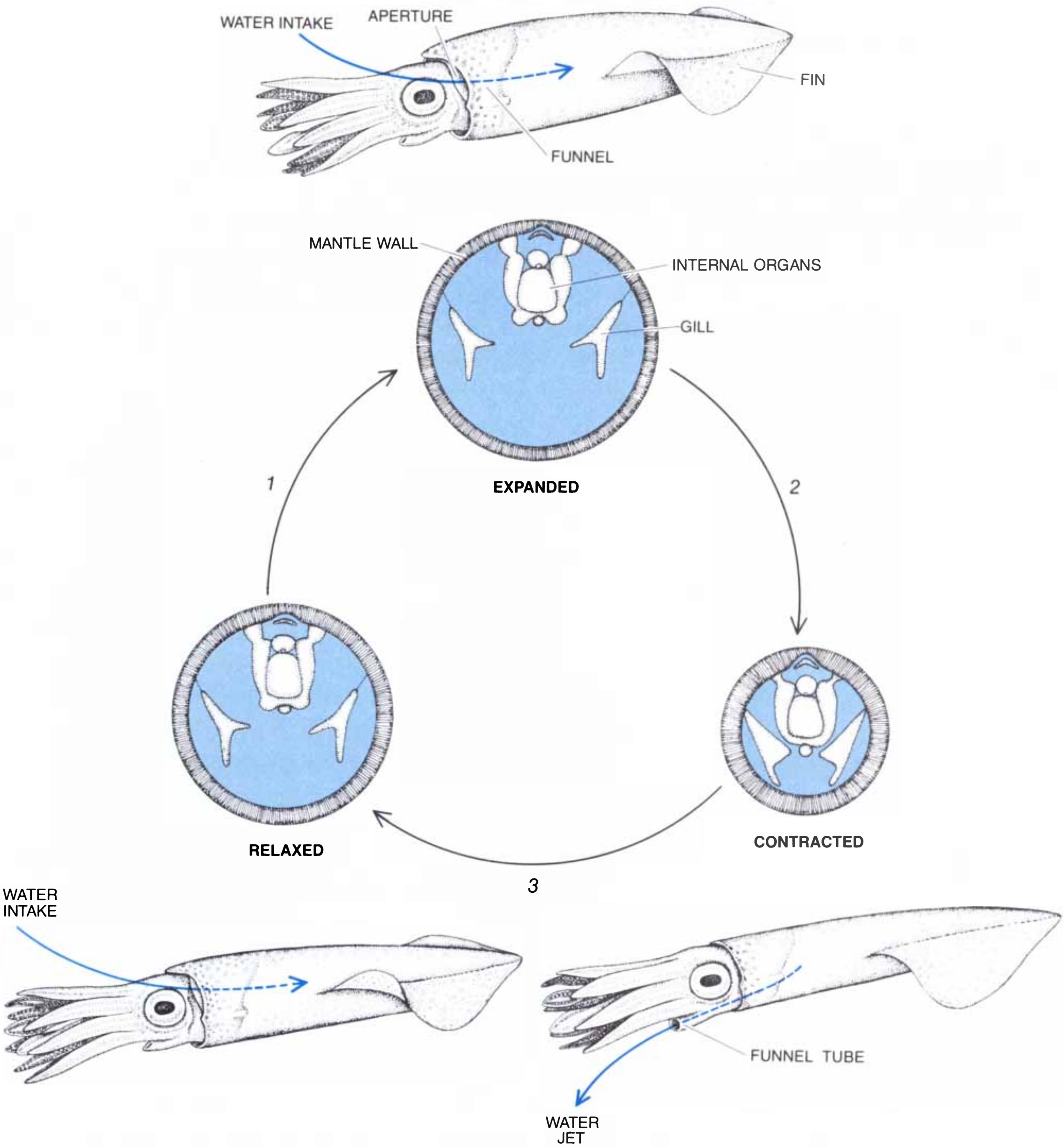
In most hydrostatic organisms the working fluid is the water or blood that fills the cylindrical body cavity, and it makes up a large portion of the animal's mass. As a result these organisms are highly flexible—think of an earthworm—but because the fluid itself generates no mechanical power, they usually are not capable of rapid movement. In squids, however, the working fluid is not the water in the mantle cavity, which is expelled in the jet, but intra- and intercellular water in the muscle tissue of the body wall. Whereas an earthworm's hydrostatic system consists of a muscular wall compressing a relatively large mass of inert fluid, the squid's is essentially the muscle itself, which means its ratio of power output to mass is potentially very high. Structurally the squid is uniquely designed to exploit the potential of this musculo-hydroskeleton to function as a high-speed swimming machine.

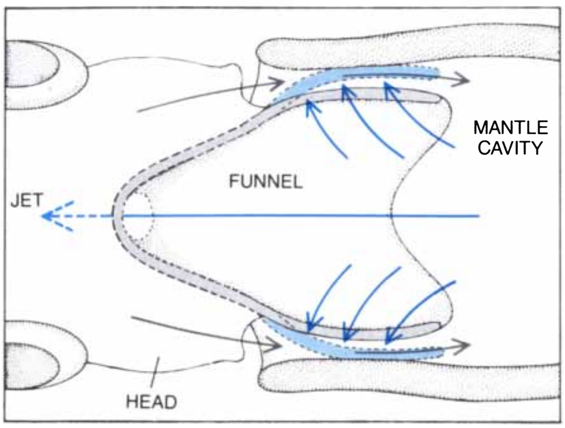
Squids are not the only animals that swim by jet propulsion. Their close relatives the octopuses do too. Although the mechanism is basically the same, the differences are instructive, and they are reflected in the animals' differing behavior patterns. Octopuses are not active swimmers; they spend most of their time creeping into and around crevices and rocks on the ocean floor.

The mantle of an octopus is a cylindrical bag formed by three independent, orthogonal (mutually perpendicular) sets of muscles: longitudinal muscles running parallel to the axis of the bag on both its inner and outer surfaces; rings of circular muscle running around the bag, and radial muscles running through the mantle wall, perpendicular to the axis. When the three muscle groups contract in unison, the muscle-tissue water is pressurized and the mantle is stiffened. If one group contracts more forcefully than the others do, the mantle changes shape.

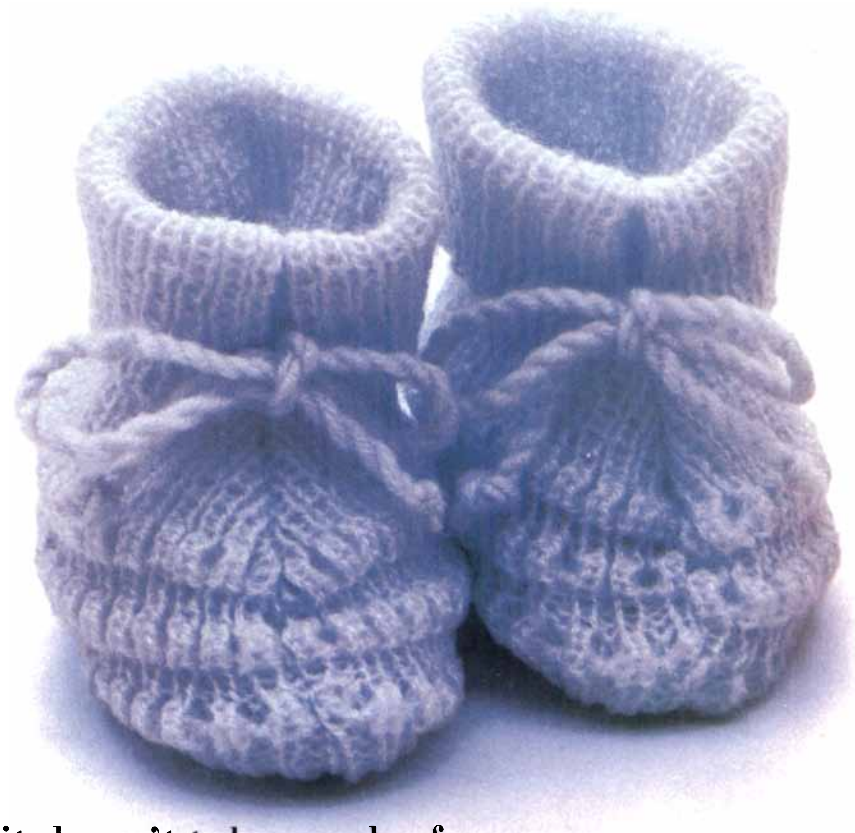
The jet is powered primarily by the contraction of the circular muscles, which decreases the diameter of the mantle and expels water from the cavity. Since muscles can do mechanical work only when they are contracting, an antagonistic muscle group must then expand the mantle so that it can refill with water. In octopuses this group consists of the radial muscles, and the antagonism is hydrostatic. When the circular muscles shorten during the jet phase, they also get wider, because the muscle volume must remain constant. The octopus simultaneously tenses its longitudinal muscles, preventing the mantle from getting longer, and so the wall gets thicker and thereby stretches the radial muscles. At the end of the jet phase the radial muscles can then contract, making the mantle wall thinner, increasing the diameter of the cavity and enabling it to refill with water. This also stretches the circular muscles and prepares them for the next cycle.

The jet mechanism in squids is similar, but with a crucial difference: as Stephen A. Wainwright and Diana Valiela of Duke University have dem-




**ESCAPE-JET CYCLE** of the squid *Loligo opalescens* is shown in side and cross-sectional views. The cycle begins with hyperinflation (*1*), in which the outside diameter of the mantle increases by approximately 10 percent compared with the relaxed state and the volume of the internal cavity rises by about 22 percent. Water flows into the cavity through large openings on each side of the head, passing over the surface of the funnel (*horizontal section at left*), a tissue structure that projects back from the head. After expansion reaches its peak, the mantle contracts to about 75 percent of its relaxed diameter (*2*). Pressure in the cavity rises sharply, forcing the funnel against the mantle wall and sealing the water inlets. Nearly all the water (equivalent to about 60 percent of the relaxed mantle volume) is expelled through the narrow funnel tube in a powerful jet. The mantle then refills to the relaxed level (*3*); further contraction might damage soft internal organs. A single escape-jet cycle lasts for about a second and may be repeated six to 10 times. When the squid is swimmming slowly, its mantle contracts, without hyperinflating, to about 90 percent of its relaxed diameter. *L. opalescens,* a common item in North American seafood markets, is typically 15 to 20 centimeters (six to eight inches) long.



In the past ten years, IBM has supported more than 90 programs designed to strengthen women's skills in these and other areas. This support includes small grants for pre-college programs in engineering, major grants for science programs at leading women's colleges, and grants for doctoral fellowships in physics, computer science, mathematics, chemistry, engineering, and materials science.

We intend to continue supporting programs like these.

Because we all have a lot to gain with men and women on equal footing. IBM®

onstrated, the longitudinal muscles in the octopus are replaced on the inner and outer surfaces of the squid mantle by layers of collagen, called tunics. Collagen is a protein fiber found in nearly all animals; in vertebrates it is the principal component of tendons and ligaments. The fibers in the squid's collagen layers are wrapped in crisscrossing helixes around the bullet-shaped mantle. (The pattern resembles that of a bias-belted radial tire.) The tunics are stiff and much stronger than the muscles they replace. They have two important functions.

First, like longitudinal muscles, the tunics prevent the mantle from stretching when the circular or radial muscles contract. But whereas the octopus must expend metabolic energy to develop tension in the longitudinal muscles, the squid need not do so because the tunics are longitudinally rigid. Furthermore, because the tunics are strong they can be thin. More of the mantle can therefore be taken up by the circular muscles that actually power the jet. The squid mantle musculature consists of alternating bands of circular and radial muscle; the circular muscle occupies about 90 percent of the total volume. The inner and outer tunics combined typically account for only 1 percent of the thickness of the wall, compared with the 15 to 20 percent taken up by the longitudinal muscles in the octopus. Thus virtually all the mass of the squid mantle is devoted to propulsion.

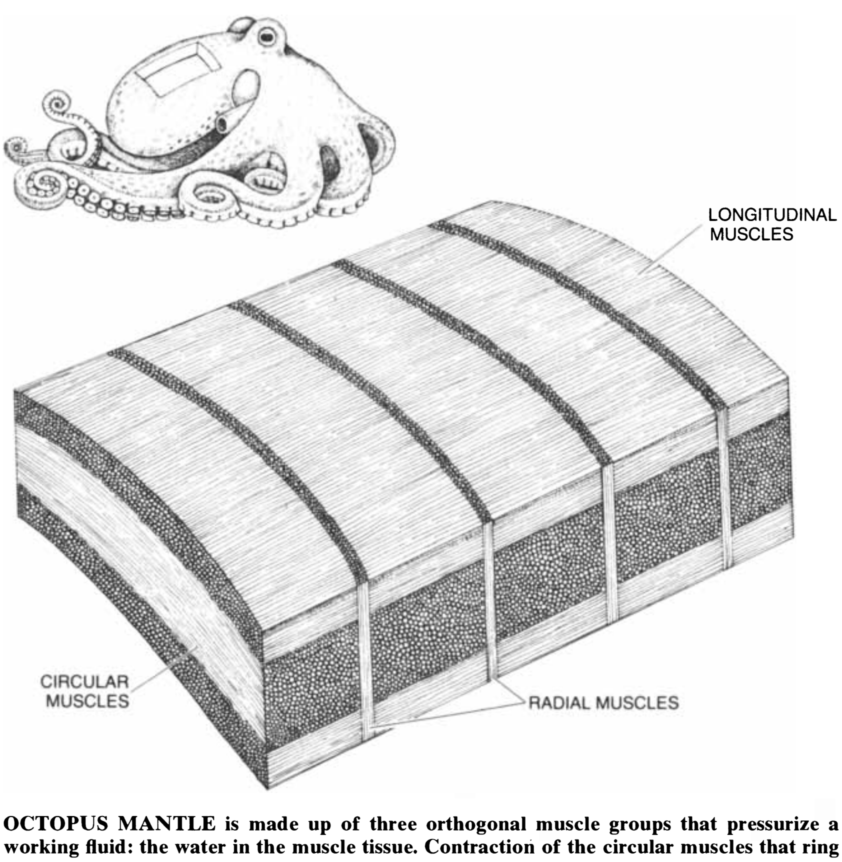
Second, the tunics provide a solid insertion surface for the radial muscles. In comparison, the insertion membranes on the inner and outer surfaces of the octopus mantle are relatively flimsy. The robust collagen tunics enable the radial muscles of the squid to contract much more forcefully after jetting, thereby refilling the mantle cavity faster. The dual role of the tunics transforms the squid mantle into a structure specialized for jetting and makes squids better swimmers than octopuses. On the other hand, the rigidity of the collagen layers prevents squids from bending easily, and so octopuses are better at oozing into crannies on the ocean floor.

The firm anchoring of the squid's radial muscles in the tunics turns out to be essential in the sequence of accelerative bursts that characterize escape jetting. At the beginning of an escape-jet cycle the radial muscles contract powerfully, and so the mantle wall becomes thinner. Since the total volume of muscle must stay the same and the mantle cannot get longer, its external diameter increases by about 5 to 10 percent over the relaxed state. This expansion, called hyperinflation, allows a maximum amount of water to be drawn into the mantle cavity, priming the animal for jetting.

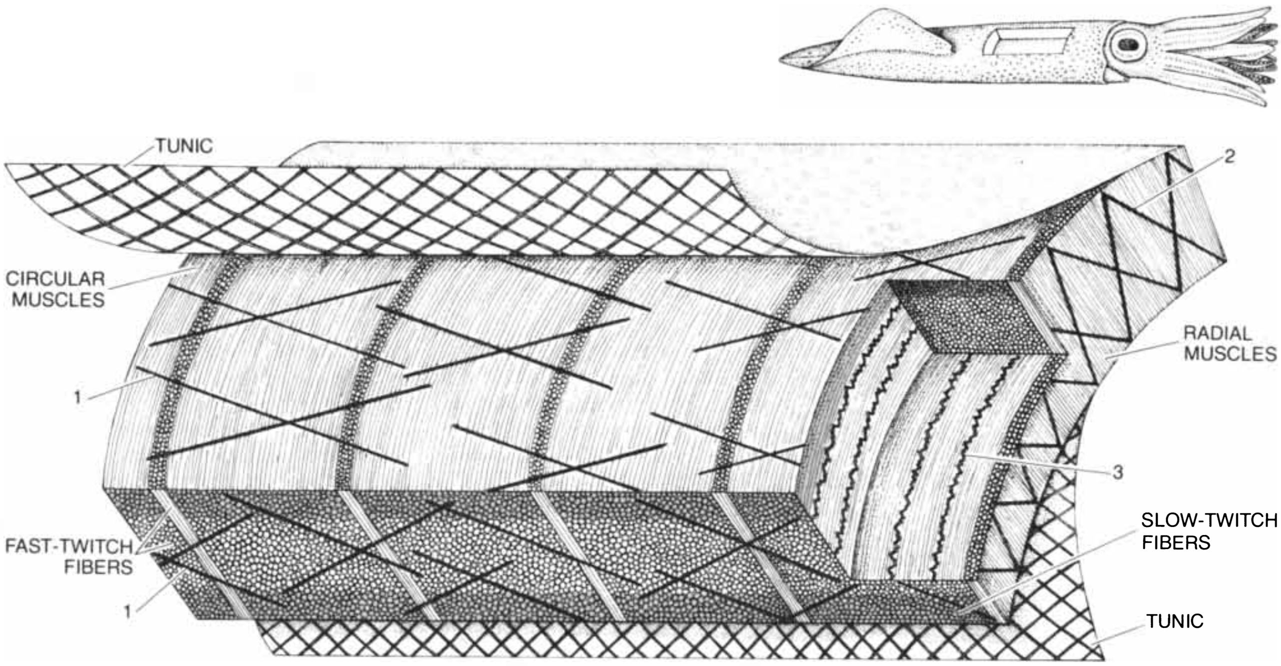
During the jet phase the contraction of the circular muscles raises the pressure in the mantle cavity to about 25 kilopascals. (Normal human systolic blood pressure is about 16 kilopascals.) The external diameter is reduced by some 30 percent, the mantle wall thickens by as much as half and the internal diameter drops by roughly 40 percent. Nearly all the water is expelled from the mantle cavity. Refilling then begins immediately; during this process pressure in the mantle cavity becomes negative, falling to about −1 kilopascal. The cycle is usually repeated six to 10 times until the animal has escaped.

The amounts of water moved in emptying the mantle and in refilling it are the same, and so the difference in pressure between the two stages indicates that the squid does much more work during the jet phase. It does so by forcing water out of the mantle cavity at high velocity through a single narrow funnel tube, whereas the water enters the cavity at low velocity through two large openings on each side of the head. The funnel, which projects back from the head into the cavity, where its wide end attaches to the mantle wall, directs the one-way flow of water. Flap valves on its dorsal surface allow water to flow into the mantle during refilling, but during the jet phase positive pressure in the cavity forces the flaps against the inner surface of the mantle wall, closing the valves and channeling the water out through the funnel tube. The tube is flexible, and the squid can swim backward, forward and in virtually any other direction by simply pointing the water jet in the opposite direction.

The jet is versatile and powerful, but it is not particularly efficient. Whereas the undulatory or oscillatory motion of a swimming fish accelerates a relatively large mass of water, a squid can move no more water than will fit inside its mantle cavity. Studies of fish that swim by tail motion suggest they achieve hydrodynamic efficiencies as high as 85 percent at optimum speeds. To achieve the same thrust as a fish of equal size, however, a squid must impart a higher velocity to a smaller mass of fluid. The higher the velocity,




**OCTOPUS MANTLE is made up of three orthogonal muscle groups that pressurize a working fluid: the water in the muscle tissue. Contraction of the circular muscles that ring the mantle bag shrinks the bag and powers the jet. At the same time tension in the longitudinal muscles prevents the mantle from lengthening, and so the wall gets thicker, thereby stretching the radial muscles. Contraction of the radial muscles thins the wall and causes the bag to refill with water. The octopus can bend its mantle and creep around rocks and into small cracks by contracting one of the two longitudinal muscle layers more than the other.**




**SQUID MANTLE has radial and circular muscles like those of an octopus, but the longitudinal muscles are replaced by stiff collagen tunics. Without expending muscle energy, the tunics prevent the mantle from lengthening when either of the muscle groups contracts. They also provide solid insertion surfaces for the radial muscles. The radial muscles are composed entirely of fast-twitch, glycolytic fibers, suggesting they contract only during escape swimming to power hyperinflation. The circular muscles are made up of a thick central layer of fast-twitch fibers and two thin layers of slow-twitch, aerobic fibers that drive slow swimming. Three sets of collagen fibers run through the muscles. The fibers in two of these systems (*1, 2*) are stretched when the circular muscles contract and thicken the mantle wall during the jet phase; the elastic energy they store is released when the mantle refills. Fibers in the third set (*3*) are stretched when the radial muscles contract during hyperinflation, and their recoil helps to expel a water jet from the mantle cavity.**

the greater the amount of turbulence in the accelerated water. This form of friction reduces the efficiency with which squids can convert muscle power into useful hydrodynamic thrust.

Accurate data on the hydrodynamic efficiency of squid jetting are not available, but measurements of the metabolic cost associated with jetting indicate how inefficient it can be. Ronald K. O'Dor of Dalhousie University in Nova Scotia recently measured the rate of oxygen consumption of squids swimming steadily in a respirometer and compared the results with those obtained from salmon of similar size. The metabolic cost of slow, continuous swimming, expressed as energy per unit of body weight per unit of distance traveled, was four times greater for the squids than it was for the salmon.

The exceptional respiratory capacity of squids helps to compensate for the inefficiency of jetting; the undulatory motion of their fins also provides a second source of power that may improve their overall hydrodynamic efficiency somewhat, particularly at low swimming speeds. Clearly, however, squids are at a disadvantage with respect to fish in continuous swimming. O'Dor has concluded that their ability to compete successfully must be attributed in large part to the accelerative power and maneuverability they show during burst or escape swimming.

The bulk of the muscle in the squid mantle seems to be active only during such high-speed bursts. In general, animals have two types of muscle fibers: "fast twitch" fibers, which power sudden, rapid movements, and "slow twitch" fibers, which power slower, more sustained motion. Fast-twitch fibers are able to contract rapidly because they have an enzyme (myosin-ATPase) that allows the breakdown of adenosine triphosphate (ATP) and the resulting conversion of chemical energy into mechanical power to proceed at a high rate. The ATP is created primarily by glycolysis, a metabolic pathway that is more readily activated than respiration. Fast-twitch fibers contain few mitochondria (the organelles in which respiration takes place); instead they are densely packed with myofilaments, the contractile elements of muscle cells. Consequently they can develop large tensions. Their poor aerobic capacity, however, causes them to fatigue quickly. Slow-twitch fibers have the opposite characteristics: their myosin-ATPase typically has a lower catalytic velocity, which makes them efficient at lower contraction velocities; their metabolism is aerobic, and so they are rich in mitochondria and resist fatigue; they contain fewer myofilaments, and so their contractions are less forceful.

T. P. Mommsen and his colleagues at the University of British Columbia and Quentin Bone and his group at the Marine Biological Association in Plymouth, England, have shown that these two distinct fiber types are segregated into three layers in the circular muscles of the squid mantle. On the basis of electron microscopy and enzyme analysis the workers identified two relatively thin layers just under the inner and outer tunics as slow-twitch muscle. A thick central layer, usually making up more than 80 percent of the total circular-muscle volume, consists of fast-twitch fibers. Presumably this glycolytic central layer contracts only during burst or escape swimming, and the aerobic, slow-twitch muscle layers are used for continuous swimming at lower speeds.

Moreover, all the fibers in the radial-muscle bands were found to be of the fast-twitch type. Assuming such fibers are active only during escape jetting, this rather unexpected finding suggested that the radial muscles have no role in refilling the mantle when the squid is swimming slowly. During slow swimming the mantle returns only to its resting diameter and does not hyperinflate. Still, something must drive the refilling. These considerations led us and other workers to investigate what may be the most interesting feature of the squid mantle: its elasticity.

Along with our colleagues John D. Steeves and Anthony D. Harman we have inserted electrodes into the mantle muscle of captive squids and recorded the electrical activity caused by nerve signals transmitted to the muscle during escape jetting. This procedure, called electromyography (EMG), enables us to determine the precise times at which muscles are contracting. By correlating the EMG traces with observations of mantle diameter we can infer whether radial or circular muscles are being activated.

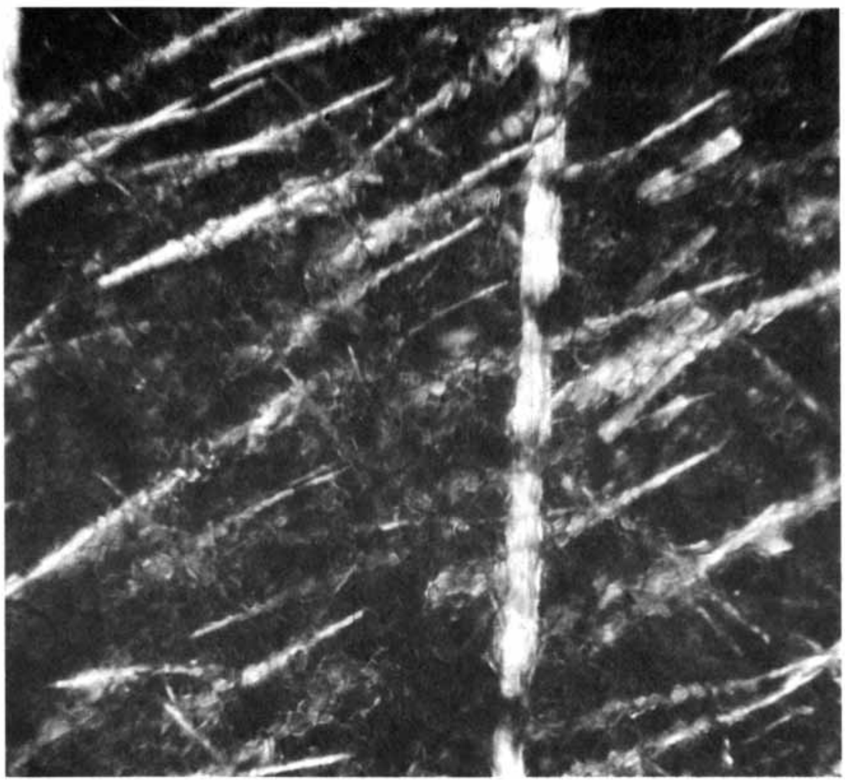
EMG peaks are always seen during hyperinflation and during the jet phase, indicating that muscle contractions power these parts of the cycle. Immediately following the jet, however, as the mantle is beginning to refill, there is often very little electrical activity. This observation tends to confirm our suspicion that an electrically silent, elastic mechanism is helping to expand the mantle. Furthermore, a detailed comparison of the EMG traces with pressure levels in the mantle cavity during jetting shows that pressure begins rising about 50 milliseconds before the first circular muscle activity is recorded, reaching about 30 percent of its peak before the muscles begin to contract. Thus an elastic mechanism also seems to contribute to the jet, at least during escape swimming. The squid mantle must contain some kind of spring that enables it to bounce back after a muscle group contracts.

Robert E. Shadwick and one of us (Gosline) have shown that the mantle does indeed have tissue springs in the form of collagen fibers laced through the muscle. These intermuscular fibers, first identified by the workers at Duke and the Marine Biological Association, are arranged in three sets. Fibers in the first set spiral through the mantle wall in a primarily longitudinal direction and attach to the inner and outer tunics. The second set of fibers also terminate in the tunics, but they are confined to the radial-muscle bands and so are oriented radially. The third set parallel the circular muscles and are not attached to the tunics.

Although the layers of collagen that make up the tunics are thick and are not extended during the jet cycle, the dispersed intermuscular fibers are very thin and are stretched reversibly by about 10 percent. In this process they can store and release large amounts of elastic energy. When the circular muscles contract during the jet phase, thereby thickening the mantle wall, they stretch the first two sets of fibers. By compressing mantle tissue in a way that simulates the action of the circular muscles, we have been able to measure the mechanical properties of these tissue springs and have found that they store energy efficiently, releasing 75 to 80 percent of it when they relax. Energy from the springs is then available to help expand the mantle during refilling. Similarly, the third set of collagen fibers are stretched when the radial muscles contract during hyperinflation. The energy thus stored is released when the radial muscles relax at the start of the jet phase. This sequence of events undoubtedly explains why pressure in the mantle cavity rises rapidly even before the activation of the circular muscles.

Our analysis of the role of the tissue springs immediately raises a perplexing question: since the energy they store comes directly from the contraction of the mantle muscles, is it not being diverted from the generation of useful hydrodynamic thrust? To answer the question we must consider both the jet cycle and the mechanical properties of the collagen fibers in more detail. It turns out that the fiber sets are constructed in such a way that they store energy only at times in the jet cycle when the muscles cannot apply their full output to hydrodynamic work. The fibers release energy when it can make a useful contribution.

From measurements of mantle diameter and of pressure in the mantle cavity we have calculated the hydrodynamic output of the jet, which is equal to the volume of water expelled times the pressure, at each stage of the cycle. During the first half of the jet phase pressure is high and the mantle diameter decreases rapidly. Most of the hydrodynamic thrust developed by the circular muscles is generated in this period. In the second half of the jet, however, hydrodynamic output drops quickly to zero even though the mantle continues to contract. The drop is a consequence of the mantle's cylindrical shape: the volume of a cylinder is proportional to the square of its radius, and so successive increments of contraction expel progressively smaller volumes of water. The resulting reduction in outflow velocity also causes pressure in the mantle to fall. Hence toward the end of the jet the circular muscles are no longer able to apply their full output to the generation of



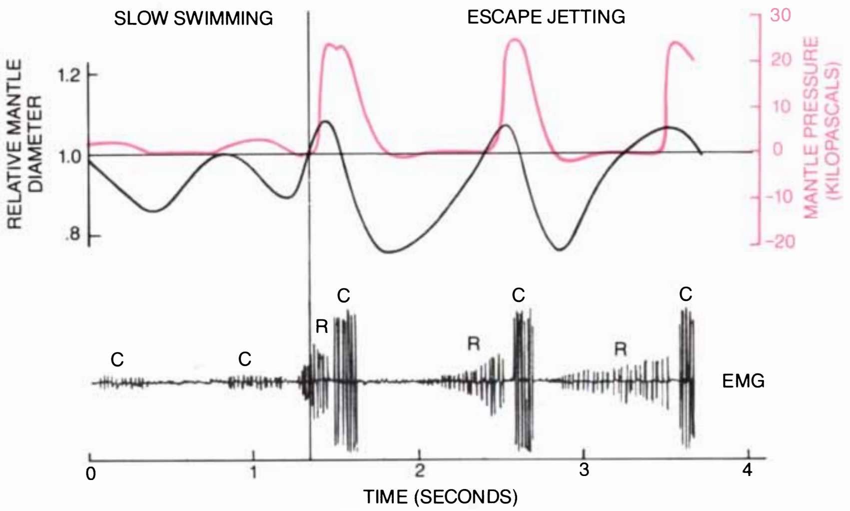
**INTERMUSCULAR COLLAGEN FIBERS are visible in this micrograph of a longitudinal section through the squid mantle wall. The cut is parallel to the front plane of the drawing on the preceding page: the light vertical band is a radial muscle and the dark areas are circular muscles running out of the plane of the micrograph. The slanted horizontal streaks are collagen fibers belonging to the first set; they spiral in a generally longitudinal direction through the mantle wall and attach to the collagen tunics on its inner and outer surfaces.**

thrust. One might expect them to relax.

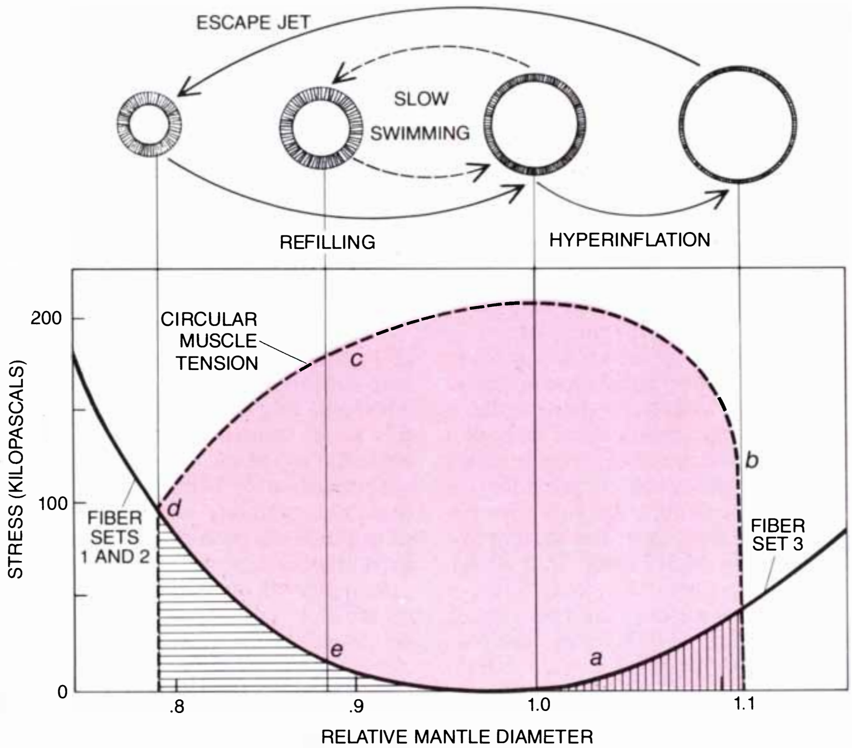
And yet they do not; it is precisely in this period that energy is stored in the first two sets of tissue springs. Our compression tests showed that the mantle deforms easily at the beginning of the jet phase, when hydrodynamic output is high. At this stage the collagen fibers absorb very little of the muscle output. Toward the end of an escape jet, however, as the level of compression approaches 20 percent, the mantle tissue becomes as stiff as solid rubber. The stiffness comes not from the relatively fluid muscle but from the collagen fibers, which, although they make up only .5 percent of the mantle volume, are about 1,000 times as stiff as the muscle. To compress the mantle tissue by 20 percent the circular muscles must maintain their tension level at about half its maximum value as the jet tapers and hydrodynamic output falls to zero. This additional muscle output is stored in the first two sets of fibers and is available to power the early stages of refilling, when hydrodynamic energy requirements are high.

Although we have not tested mantle tissue under conditions of expansion, it is probable that the third set of fibers store energy from the radial muscles in a similar way during hyperinflation. The hydrodynamic work needed for hyperinflation is essentially zero, because water enters the mantle cavity through large openings at low velocity and negligible pressure. EMG recordings indicate clearly, however, that the radial muscles are very active at this stage. We suspect that nearly all their output is stored in the collagen fibers. The elastic energy is then released to increase the power output at the beginning of the jet, again when energy requirements are high. Together the three sets of fibers enable the squid mantle to generate greater thrusts than it could if it were made only of muscle, and they do so without adding significantly to its overall mass.

Squids owe their power as swimmers to a skeleton in which fluid dead weight has been replaced by muscle and in which muscle is judiciously stiffened by collagen. Their extraordinary swimming performance in turn accounts for their ability to compete with fish. How successful are they? One indication is their remarkable prevalence in the ocean, where they account for a large fraction of the total biomass. Reliable data on squid numbers are not available, but it has been estimated that sperm whales alone eat about 100 million tons of squid per year. In comparison, the total catch of fish by the world's commercial fishermen amounts to only 70 million tons.




**ELECTROMYOGRAPHIC RECORDINGS (EMG) from the mantle muscles of captive squids show peaks corresponding to radial-muscle contractions (*R*) during hyperinflation, when mantle diameter (*black curve*) is increasing beyond the relaxed level; the strong peaks during the subsequent jet phase correspond to circular-muscle contractions (*C*). Virtually no muscle activity is evident immediately after the first escape jet, suggesting that an electrically silent, elastic recoil mechanism is powering the refilling of the mantle. Squids swimming freely, however, probably show a pattern of muscle activity closer to that of the second jet cycle, in which radial-muscle contractions accelerate refilling. In slow swimming the radial muscles are not active; the mantle does not hyperinflate, and pressure levels in the mantle cavity (*colored curve*) are much lower than those characteristic of escape jetting.**




**ELASTIC ENERGY stored in collagen fibers supplements muscle power in squid jetting. The solid curves show the levels of stress required to stretch the three sets of fibers by the amounts associated with various mantle diameters. During hyperinflation (*a*) the contraction of the radial muscles expands the mantle diameter by about 10 percent and stretches the fibers that parallel the circular muscles. Elastic energy (*vertical hatching*) is stored in the fibers and is then released as circular-muscle tension rises sharply at the beginning of the jet phase (*b*). At first nearly all the circular-muscle energy is converted into hydrodynamic work (*colored area*), but as the mantle contracts to less than its resting diameter (*c*) an increasing portion of the output is devoted to stretching the two sets of fibers that traverse the mantle wall. Contraction stops (*d*) when muscle tension is no longer enough to stretch the fibers; energy stored in them (*horizontal hatching*) helps to power refilling (*e*).**

# The Crossbow

*This formidable weapon, invented 2,400 years ago, became popular in the 11th century. For 500 years, until the advent of efficient firearms, it was preponderant in defensive situations*

by Vernard Foley, George Palmer and Werner Soedel

From the 11th century until the advent of efficient firearms some 500 years later the crossbow was a formidable weapon of warfare. It served mainly in the defense of fortified or protected places, such as castles and ships. It also made significant contributions to the understanding of materials (because of the forces the bow had to withstand) and aerodynamics (because of the peculiar flight characteristics of the bolt). The aerodynamics of its bolt and the principles of its operation inspired many of Leonardo da Vinci's excursions into physics and engineering.

The armorers and fletchers who designed the crossbow and related weapons lacked mathematics and other formal engineering knowledge. Yet tests we have carried out at Purdue University show that these craftsmen managed to achieve a high degree of aerodynamic sophistication and a good grasp of mechanical principles.

As machines go the crossbow is not complex. A bow, usually too powerful to be drawn without mechanical aid, is mounted transversely at the front of a stock of wood or metal properly called a tiller. Some means is provided for catching the string at the full-draw position and releasing it. The short arrow (the bolt) is guided either in a trough cut into the upper side of the tiller or by rests that support the bolt at each end. If the bow is sufficiently powerful, a device for drawing it will be built into the tiller or carried separately.

The design offered two advantages over the handbow. First, the crossbow achieved a greater range, so that crossbowmen could fire with impunity at handbow archers who were still too far away to respond. Second, the tiller, cocking and string-release devices partially mechanized the firing cycle, reducing the effort and skill needed to use the weapon. The claws that grasp and release the string and bolt represent one of the early attempts to mechanize some of the functions of the human hand.

A major disadvantage of the crossbow was its slow rate of fire compared with that of the handbow. Consequently the military use of crossbows was limited mainly to situations in which cover was available during the reloading process. That is why they were chiefly popular for castle garrisons, siege trains and ships.

The crossbow was invented long before it became popular. There are two main claims for the invention, one from Greece and one from China. In about 400 B.C. the Greeks developed the catapult, a machine for hurling rocks and arrows. The idea grew out of efforts to increase the power of handbows. The catapult, which resembles the crossbow, eventually became quite large, but some of the early ones seem to have been of crossbow size.

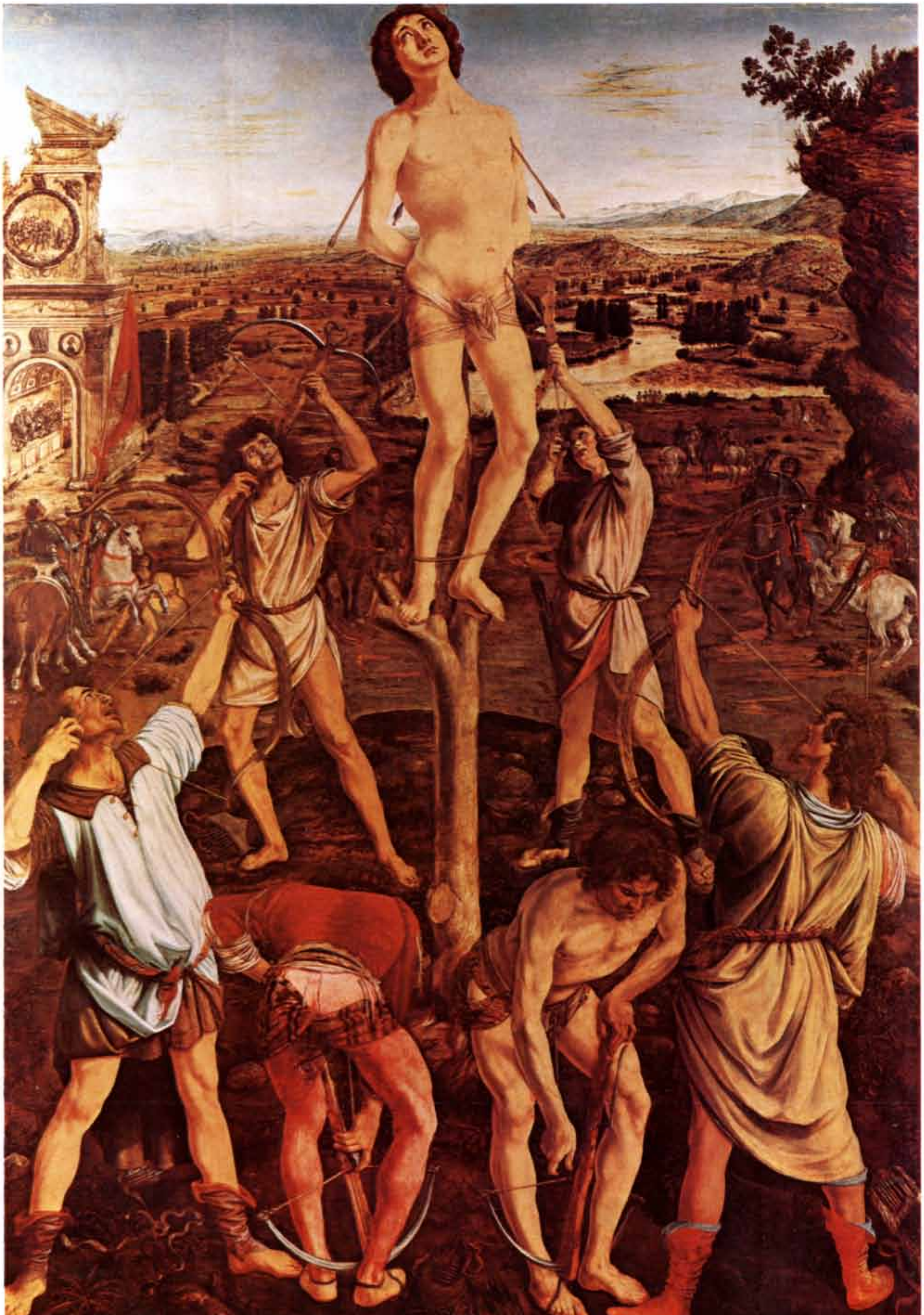
The claim of China is supported archaeologically by bronze release mechanisms made in about 200 B.C. Although the Greek claim precedes this date, written Chinese accounts place the use of the crossbow in battle at approximately 341 B.C. Other narratives, the reliability of which is harder to establish, push the story back at least another century.

It appears from the archaeological record that the use of the crossbow continued unbroken in Europe from classical times to its period of greatest popularity from the 11th century into the 16th. Two factors seem to have limited its appeal before the 11th century. One factor was that it cost more to arm troops with crossbows than it did to arm them with handbows. The other was the relative scarcity of castles. They became historically important only in the Norman period.

With the rise of the castle the crossbow became an integral part of a violent and severely elitist social revolution. Fortifications in pre-Norman times were often fairly simple affairs, designed for occasional use and intended to shelter the entire population of an area. Thus there would be plenty of firepower inside the walls when the inhabitants took refuge against a marauding band. The Normans imposed control through a small, heavily armed military minority that dominated the much larger peasant and town populations. Their castles were intended to shelter the few from the many as well as from armed and predatory members of their own caste. The superior range of the crossbow helped them to secure these refuges.

During the centuries that followed the emergence of the crossbow as a defensive weapon efforts were made to improve its power. One of the improvements may have been borrowed from the Arabs. Arab handbows were of a type called composite. Their structure merits the name, for they are important ancestors of today's composite materials. A composite bow offers clear advantages over one made from a single piece of wood. Such a bow is limited in power by the innate strength of the material. As a bow is drawn, the material on the outer (convex) side of its limb (called the back because it faces away from the archer) is placed under tension. The inner (belly) side is placed under compression. If the bow is overdrawn, splinters of wood traditionally called slithers begin to spring loose from the back, and crinkles that are called frets or crystals appear on

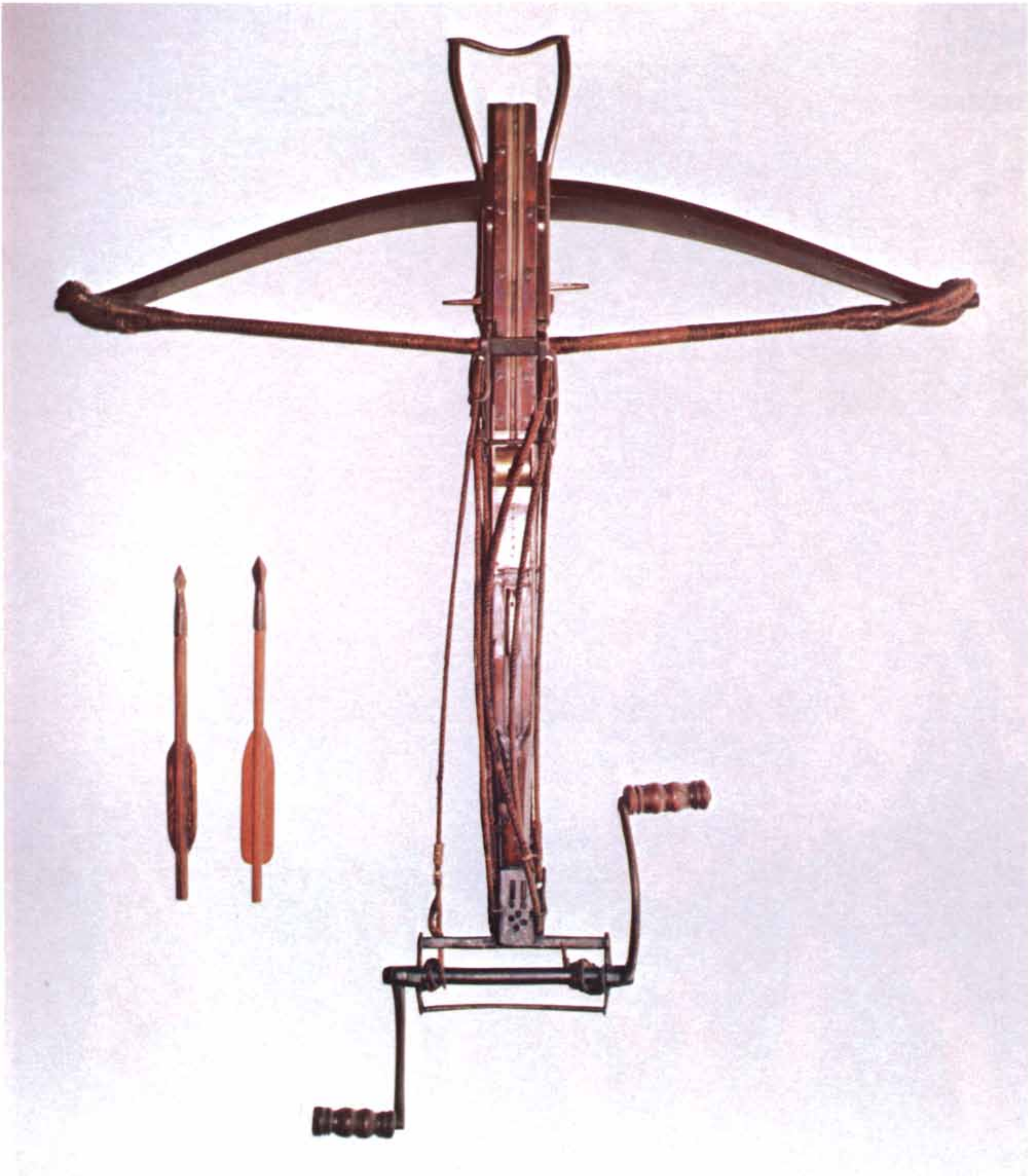
**THREE CROSSBOWS appear in this painting, *The Martyrdom of St. Sebastian,* by the 15th-century Italian artist Antonio Pollaiuolo. One archer is aiming a crossbow at the martyr and two men are cocking crossbows with the aid of stirrups, which were necessary because of the bow's strong draw weight. The painting is in the National Gallery in London.**



the belly. Usually by then the limb of the bow has become permanently bent. Further strain may cause it to snap.

In a composite bow material that is stronger in tension than wood is laid as a reinforcing layer over the back of the bow, taking some of the strain from the wood and keeping down the formation of slithers. A favorite material was animal sinew, particularly the *ligamentum nuchae:* the large, elastic bundle that runs along the spine and over the shoulders of most mammals. In tests of this material we have found that when it is suitably prepared, it has a tensile strength approaching 20 kilograms per square millimeter. That is about four times the feasible stress range of most appropriate woods.

On the belly side of the bow might be fastened a material stronger in compression than most woods. The Turks used buffalo horn, which has a maximum compression strength of about



**FRENCH MILITARY CROSSBOW of the 14th century and two of its bolts, or arrows, are on display at the museum of the U.S. Military Academy at West Point, N.Y. The bow is too strong to be drawn by hand, and so a windlass fitted over the rear of the tiller, or stock, serves for cocking. The tiller is 40 inches long; the bow is 42 inches wide; the length of the bolts is approximately 15 inches.**

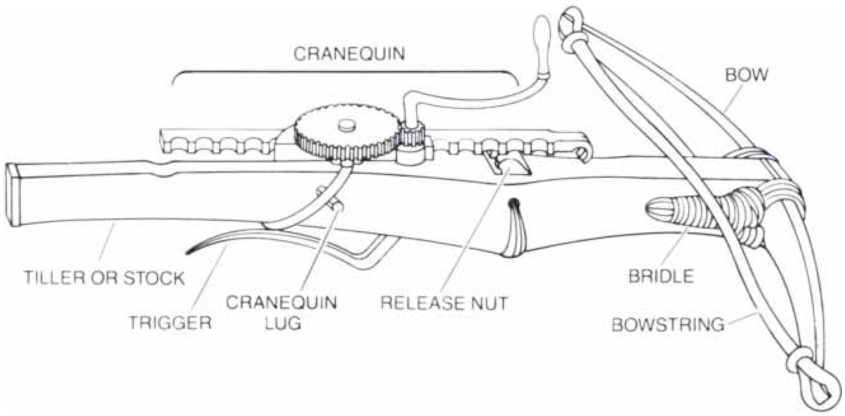
13 kilograms per square millimeter. (Wood can be almost four times weaker in compression than it is in tension.) The level of empirical sophistication reached by the makers of composite bows can also be seen in the glues they used. The favored glue was made from the skin of the roof of the mouth of the Volga sturgeon. This degree of specification testifies to an extensive tradition of experiment and selection.

Composite crossbows continued in use through the later Middle Ages and on into the Renaissance. They were lighter in weight than the steel crossbows that began to be produced in about 1400, they would shoot farther for a given draw weight and they were less likely to fail catastrophically. Such bows were fairly common in Leonardo's time. His papers indicate that he meditated on the construction of bows and derived from them fundamental insights on how materials behave under stress.

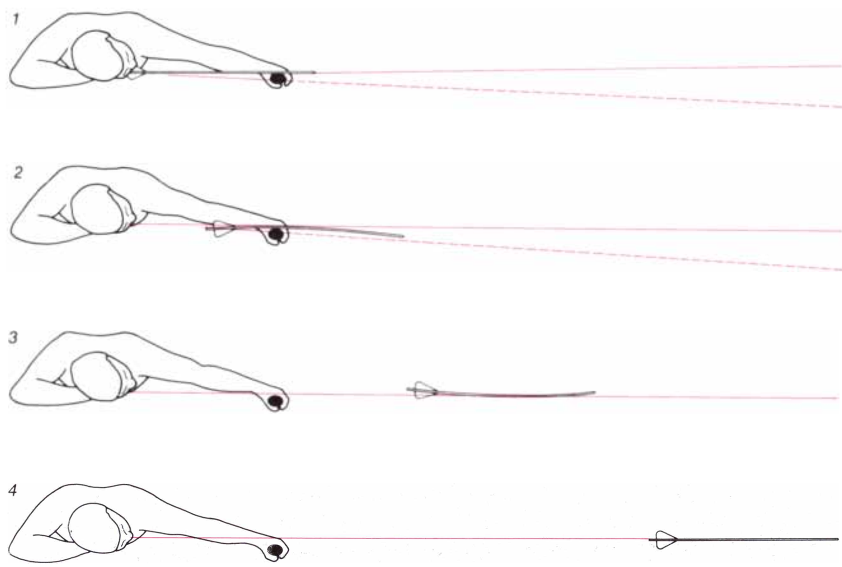
The steel medieval bow represented the zenith of crossbow construction. Its performance went unmatched until the appearance of fiberglass and other modern composites after World War II. Steel bows could be made to power levels that no organic material of the time could equal. The Victorian sportsman Ralph Payne-Gallwey, who wrote a classic treatise on the crossbow, tested a large war bow that had a draw weight of 550 kilograms (1,200 pounds) and threw an 85-gram (three-ounce) bolt 420 meters (460 yards). Egon Harmuth, a historian of the crossbow, believes bows with draw weights that were twice as high existed. The longbowmen of the time were limited in the main to bows with a draw weight of less than 45 kilograms (100 pounds). Even with special lightweight arrows they apparently achieved a range of less than 275 meters (300 yards).

As steel bows reached new force levels they encountered diminishing returns. The increased mass of the bow limb put a ceiling on the bow's acceleration capability. It was difficult to produce steel in ingots large enough to make a complete bow, and so the limbs were usually welded up from many smaller pieces. Each weld entailed an enhanced opportunity for failure and also presented something of a threat to the bowman.

The more powerful bows created a need for more powerful release mechanisms. Up to this point European releases, consisting usually of a revolving nut and a one-lever trigger, were far inferior to the Chinese mechanisms, which had an intermediate lever that enabled the bowman to fire a heavy bow with a short, crisp and light pull on the trigger. Soon after 1500 improved multiple-lever triggers began to appear in German shooting matches. Manuscripts from about a decade earlier show that Leonardo had already arrived at this design and had calculated the mechanical advantages.




**COMPONENTS OF A CROSSBOW** include the curved bow itself, the bowstring, the release nut (which held the string) and the trigger. A pull on the trigger depressed the nut, releasing the string and firing the bolt. The cranequin lug positioned the cranequin, which pulled back the bowstring. The cranequin is an early example of heavily loaded gearing.




**ARCHER'S PARADOX** accounts in part for the shortness of crossbow arrows. The paradox is depicted as it arises when an archer fires a conventional handbow arrow. As he takes aim (*1*) the arrow passes along one side of the bow. His line of aim is along the arrow. When he fires (*2*), however, the force of the string pulls the tail of the arrow toward the center of the bow. If the arrow is to stay on the line of aim, it must flex in flight (*3*). For the first few meters of flight it vibrates, but eventually it stabilizes on the flight line (*4*). The need for flexibility limits the amount of energy that can be put into the arrow. Because of the large amount of energy generated by the crossbow, it required shorter, stiffer arrows. They did not flex. Shortening the arrows also provided better aerodynamics in flight.

The crossbow bolt also improved considerably over the years. The evolution is best understood by looking first at the forces on a handbow arrow. If the archer shooting a traditional handbow is to be comfortable, the arrow must reach roughly from the center of his chest to the end of his outstretched arm. He aims the arrow by looking along the shaft, the ends of which are positioned by his hands. These two points determine the direction the arrow should take on release.

The forces that play on the arrow during its release, however, do not quite coincide with this sighting line. The released string will pull the butt of the arrow toward the center of the bow rather than toward the side. Hence if the arrow is not to be deflected from its line of aim, it must flex slightly as it is launched.

The necessary flexibility of the traditional handbow arrow sets limits on the amount of energy that can be put into one during acceleration. For example, we have found that an arrow designed for a handbow in the 20-pound draw range can be sprung far enough out of line for the shaft to break when it is fired from a crossbow with a draw of about 85 pounds.

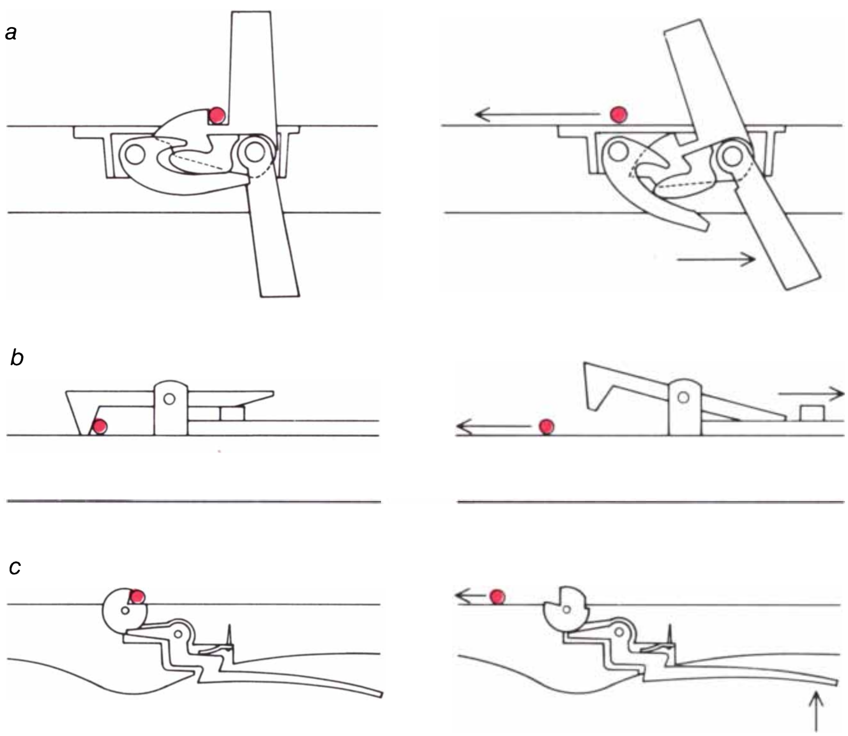
Thus arrows had to be redesigned in antiquity for use in crossbows and catapults. Because the surface of the tiller made the motion of the string coincide better with the initial line of flight and its guide arrangements replaced the archer's hands, bolts could be made shorter and stiffer. That in turn made them easier to store and transport.

The designs that emerged can be represented by two main surviving types of arrow. One of them is about half as long as a conventional handbow arrow. It flares sharply toward the rear and has vanes that by themselves are too small to stabilize the arrow well. Its butt is designed to be gripped by a release claw.

The other main type of arrow has no vanes or feathers. Its metal head constitutes about a third of the overall length, and the wood shaft has been reduced to the minimum length needed to guide the head forward through the air. It too is flared at the rear. The overall length is less than 150 millimeters (six inches).

These bolts indicate considerable aerodynamic experience on the part of the Roman innovators who first designed them. Today it is well understood that the vanes, which keep the arrow from rotating in flight, are a major source of drag on an arrow. Reducing their size would enable the arrow to fly farther, provided it did not begin to fly sideways, which would greatly increase the drag. One solution is to taper the shaft so that it is narrower at the front than at the back. If a tapered shaft begins to veer, the total air pressure will be greater on its enlarged rear segment than on its smaller front section; thus it will recover its proper alignment.

Put another way, the shaft can be regarded as having a center of pressure (the point of balance of all the aerodynamic forces acting on the shaft) behind the center of gravity. On a cylindrical arrow without vanes this point will be at about the middle of the shaft. On a tapered bolt the center of pressure is moved toward the rear by virtue of the enlargement there. Because the center of pressure is aft of the center of gravity, such an arrow is stabler than a cylindrical shaft and experiences less drag than one with vanes. Enlarging the shaft toward the rear also helps to keep the airflow pressed smoothly against the surface. In the terminology of modern aerodynamics the boundary layer is less likely to break down. Shortening the arrow shaft also helps, because airflow parallel to a cylindrical surface tends to become more turbulent as the length increases. The result is that energy-robbing turbulence is kept to a minimum along the surface of the shaft.



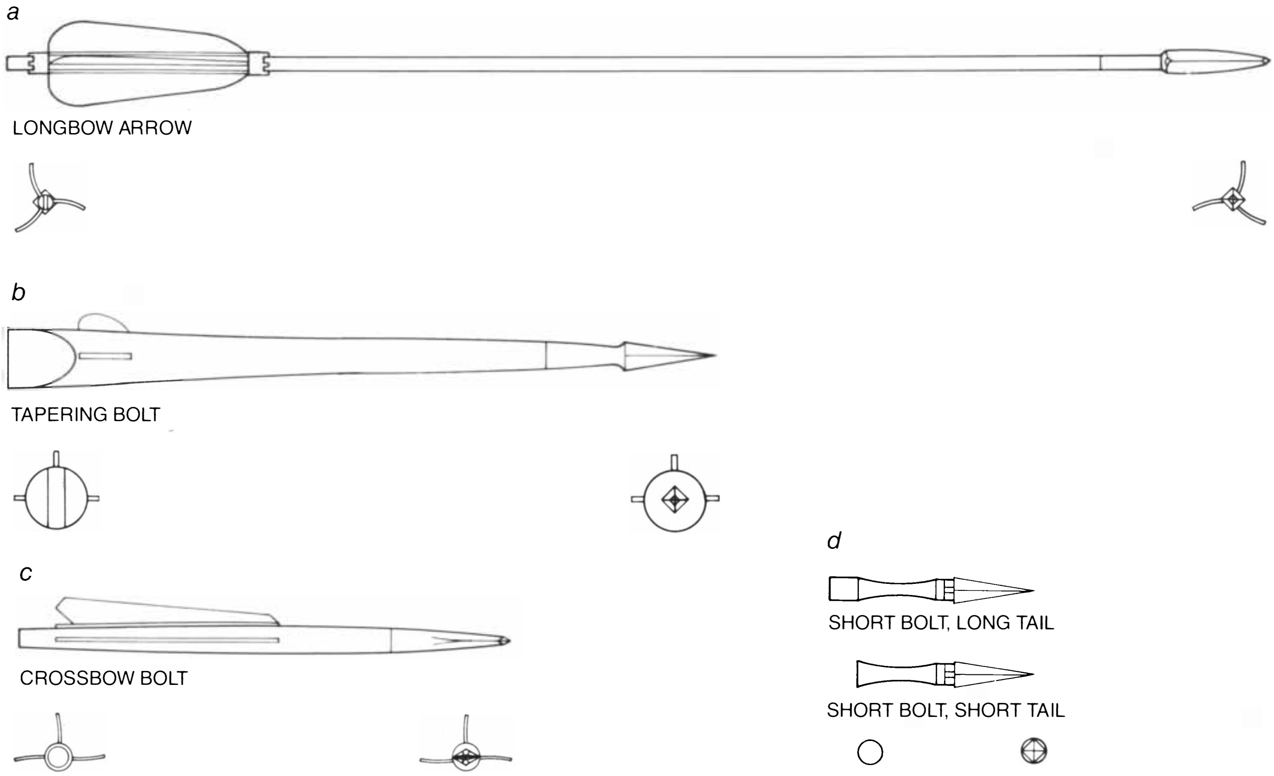
**FIRING MECHANISMS for crossbows were of several varieties. A Chinese design of some 2,000 years ago (*a*) had a claw for the string that pivoted on the same pin as the trigger. A curved intermediate lever connected the two parts and made possible a short, light pull on the trigger. The movement of the string when the trigger is pulled is shown at the right. The Western release technology first appeared in the gastraphetes (*b*), a catapult mechanism. Here the claw rises instead of falling to release the string. The standard medieval release in Europe (*c*) employed a cylindrical piece called a nut; it was supported against the force of the bow by a simple lever trigger (the key) that engaged a notch on the underside of the nut. The archer's pull on the long trigger tended to deflect the aim. Gradually systems of intermediate levers came into use to make the pull shorter and lighter.**

Another factor accounting for the superior efficiency of the tapered bolt seems to lie in the design of its tail. The tail is nocked in order for the bolt to fit between the claws of a catapult trigger. Like the tapering, the nocking helps to smooth the flow of air past the stern of the projectile, reducing the formation of a turbulent and energy-dissipating wake.

There is no reason to assume the technical experts of the time knew about pressure or the details of airflow and drag. These ideas began to emerge only with Leonardo. Doubtless the ancient bolts were designed by trial and error mixed with logical deductions. Probably the guiding factors were maximum range and force of impact.

Nevertheless, the craftsmen of the time achieved substantial improvements in the design of projectiles. Wind-tunnel tests that we have made at the Purdue University Aerospace Sciences Laboratory substantiate this conclusion. We tested a typical medieval war-handbow arrow, a medieval quarrel (a crossbow bolt) and the two known designs for ancient catapult bolts. Our findings should be interpreted with some caution, because the sizes of these projectiles, particularly the smallest one, approached the sensitivity limits of the measuring apparatus. Even within the limits of our analysis, several interesting tentative conclusions emerged. First, the smallest bolt, which appears to be fairly complete except for some deterioration at the tail, was probably stable at all the flight angles that could reasonably be expected for it in use.

Second, our comparison of the ratio of drag to weight among the four projectiles reveals that the handbow arrow is substantially inferior to the others. The weight can be regarded as a



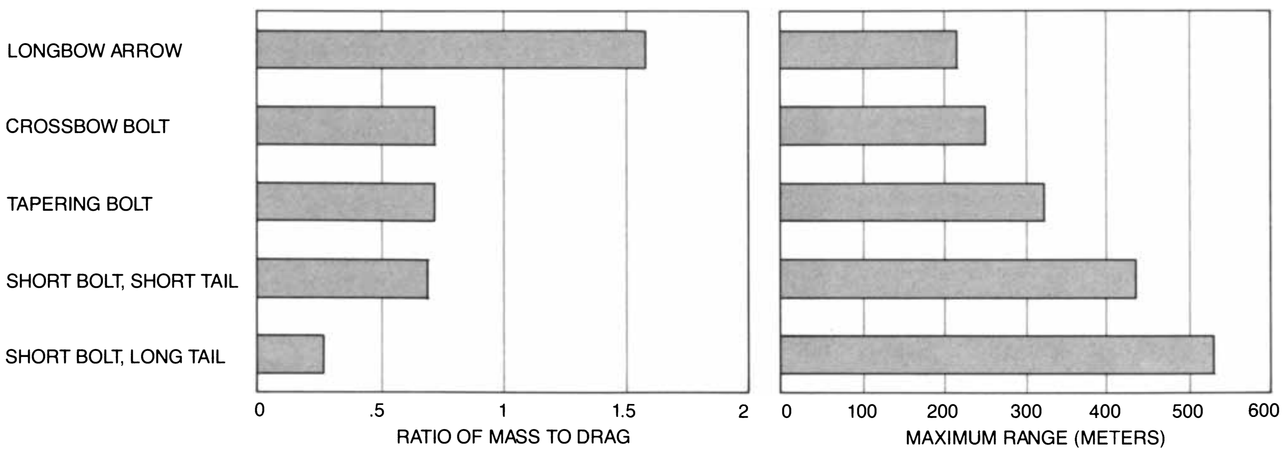

**PROJECTILES FOR BOWS include a typical arrow for a war longbow, a variety of handbow (*a*); a tapered Roman bolt launched from a catapult resembling a crossbow (*b*); a typical medieval crossbow bolt (*c*), and two variations on another Roman catapult bolt made for a smaller machine (*d*). Each upper view is from the side; below are end views of the corresponding part of the projectile.**

measure of the projectile's ability to store energy. If all the projectiles were launched at the same velocity, their weight would determine their energy at the beginning of flight. Their drag determines the rate at which energy will be lost. A low drag-to-weight ratio indicates a likelihood that the projectile will achieve a long range.

The drag-to-weight ratio of the handbow projectile is about twice the ratio of the other missiles. It appears that once the design constraints on handbow arrows could be disposed of the ancient and medieval designers were able to arrive at a new level of optimal design. The design was so well suited to the materials then available that it was not improved on during the remainder of the period in which archery was the chief mode of warfare.

Behind all this mechanical development lay a strong need for crossbows and their ammunition. Often the bulk of the peacetime garrison of a castle consisted of crossbowmen. At a heavily defended outpost, such as the English port of Calais on the coast of France, as many as 53,000 bolts were kept on hand. The authorities there habitually bought bolts in lots of 10,000 or 20,000. From 1223 to 1293 the de Malemort family of the Forest of Dean in England produced nearly a million bolts.

The result, as one might expect, was a trend toward mass production that far antedates the Industrial Revolu-




**WIND-TUNNEL TESTS of the five types of projectile shown in the top illustration on this page were made by the authors at the Purdue University Aerospace Sciences Laboratory. The results are charted here. The estimates were computed by Wade H. Hickam for an initial velocity of 80 meters per second—probably too high for the longbow, but it serves as a common speed for comparison.**

tion. One example is a device consisting of a pair of wood blocks screwed together to form a clamp; the blocks were grooved to accept an arrow shaft. The vanes of the arrow protruded through slots and rested on a metal plate that served as a guide in trimming them to the proper size and symmetry. This operation of fitting and trimming the vanes is called fletching. A medieval occupation important enough to leave behind such a fairly common surname as Fletcher began to be handed over to mechanical aid.

Another device is a planing machine that probably served both to round off the bolt shafts and to cut the grooves that the vanes fitted into. Wood shafts of small diameter could not easily be made true in the lathes of the time because they flexed out of line when the cutting tool was applied. In the planing machine the cutting iron is held in a wood block with two opposing handles. The block slides along the body of a clamp that supports the arrow shaft rigidly. The iron will cut until its block rests on the top surface of the clamp body. Thus the device provides automatic control over the depth and direction of the cut, and the bolts can be held to almost identical size.

After Leonardo's time the crossbow began to decline in popularity as firearms came into wide use. Crossbows continued to be used at sea, however. They survived there because they did not have the ignition problems that plagued early firearms and because the bulwarks offered protection for reloading. Heavier versions also continued to be used for whaling. On land firearms gradually replaced crossbows for hunting. An exception is a version of the crossbow that shot a stone or a ball. This weapon, used for hunting small game, survived into the 19th century. That this pellet or bullet bow apparently owes some of its features to firearms represents a reversal of the evolutionary relation between the two kinds of weapon. Such features of firearms as stocks, hair triggers and adjustable sights were first developed for crossbows, particularly the target bows that still survive in many parts of the world.

The development of fiberglass construction in the 20th century has resulted in the composite crossbow's seeing service again. The glass fibers offer a modern counterpart to sinew, and their plastic matrix has replaced buffalo horn. Although in the resurgence of archery the crossbow has run a far second to the handbow, it definitely has a following. The contemporary archer has at hand a crossbow vastly improved over the medieval weapon.

# THE AMATEUR SCIENTIST

## *Searching for patterns of rainfall in a storm*

by Jearl Walker

Does the rate of rainfall in a storm have a pattern or is it random? Richard W. Stimets of the University of Lowell has devised an experiment with which one can search for such patterns. Analyzing the storms that pass over the eastern end of Massachusetts, he has found that the rainfall in some of them does indeed form certain patterns, indicating that the moisture in clouds is occasionally organized.

What might cause the organization? Storms in North America are normally associated with low-pressure regions that travel approximately eastward. The low-pressure system is a great mass of air made to circulate counterclockwise by the rotation of the earth. A typical storm has two fronts, or boundaries separating air masses of different temperatures and winds. On the leading (eastern) side of the storm system is a warm front. This boundary extends from the center of the system to the east or southeast. On the northern side lies cold air. On the southern side is warm air brought up from the Gulf of Mexico by the counterclockwise circulation of the system. The warm air tends to ride up over the cool air. In the southwestern part of the system is a cold front, which is a boundary that extends from the center of the system and toward the southwest. On the southeastern side is warm air. On the northwestern side is cold air brought down from Canada by the circulation. The cold air tends to slide in under the warm air.

In a warm front warm air rises over cold air along a sloped boundary. As the moist, warm air rises it expands and cools; some of the moisture condenses to form precipitation. A cold front is formed by a mass of cold air intruding on warm air. The warm, unstable air lifted by the intrusion forms clouds and storms. The cold front travels faster than the warm front and eventually overtakes it. The combined front is said to be occluded.

I shall limit my discussion to the storms of younger systems in which the cold front is separated from the warm one. Both fronts can produce rain. Can they also produce atmospheric waves that force the rainfall into a general rhythm? Do they generate distinct concentrations of precipitation separated by some characteristic distance?

To monitor the patterns of rainfall Stimets employs a tipping-bucket rain gauge he obtained from Rainwise, Inc. (P.O. Box 443, Bar Harbor, Me. 04609). A funnel on the top directs water into one of two buckets below. When the bucket is full, it tips, losing its water and bringing the other bucket into position. The capacity of each bucket is equivalent to .01 inch of water spread over the collecting area of the funnel.

Each time a bucket tips it closes a reed switch. The switch shorts a circuit that is connected to a strip-chart recorder. The strip-chart recorder displays each tip of a bucket as a spike. One can easily determine the rainfall rate by scanning the chart: the spikes are well separated during light rain but cluster during heavy rain.

The average rate of rainfall can be computed from the chart in two ways. One way is to assume that the rate is constant during the interval between tips. The rate is therefore equal to .01 inch of rain divided by the time between tips.

One might also fit a function to the data by means of interpolation. When the rainfall rate is plotted on a graph, the interpolation yields a smoother curve than the first method does. The graph also displays peaks that are somewhat higher.

Rainfall rates can vary considerably during a storm, ranging from less than .001 inch per hour to more than four inches. What Stimets looked for was evidence of periodicity. He studied data on 30 storms from a rain gauge at Wellesley, Mass. All the storms produced at least .4 inch of rain and lasted for at least four hours. He divided much longer storms into eight-hour segments. He then fitted an interpolation function to the data so that he could compute the average rates per minute.

Next Stimets did a Fourier analysis of the average rates. This analysis reveals any sinusoidal variations in the data even when they are buried in noise. The analysis yielded information about which periods of sinusoidal functions best fit the average rates of rainfall.

A typical result for a single storm is a graph called a power spectrum. It reflects sinusoidal variations in the rainfall. The horizontal axis shows the time periods on which the Fourier analysis was based; the vertical axis shows how often the sinusoidal variations appear in the rainfall [*see top illustration on page 116*].

For convenience Stimets applied a scaling factor so that the longer periods are more condensed on the horizontal axis than the shorter periods. The graph peaks at about 160 minutes, indicating that a sinusoidal variation with a period of 160 minutes best fits some of the data. In other words, rainfall intensified approximately every 160 minutes.

The graph for a single storm may be too noisy to interpret. In order to improve the peaks representing periodic variations Stimets combined the data from many storms of the same type. He assumed that a recording station can sample four kinds of low-pressure system: a cold front, a warm front, a land cyclone and a coastal cyclone. On the basis of multiple samples of such storms Stimets has concluded that all but the coastal cyclones may show periodicity in rainfall.

In the first case the station samples the passage of a cold front. If the low-pressure system exists, it lies to the north of Massachusetts, probably in Canada. In the second case the station samples a warm front. The low-pressure center would be west of the station, somewhere near the Great Lakes. In each case Stimets considers the low-pressure center to be far enough away so that there is relatively little movement of air at the station.

The third case involves a storm pattern called a land cyclone. The low-pressure center of the system is so close to the station (within 500 kilometers) that the cyclonic wind circulation may affect the rate of rainfall. The station may sample both warm and cold fronts, usually receiving more rain from the warm front.

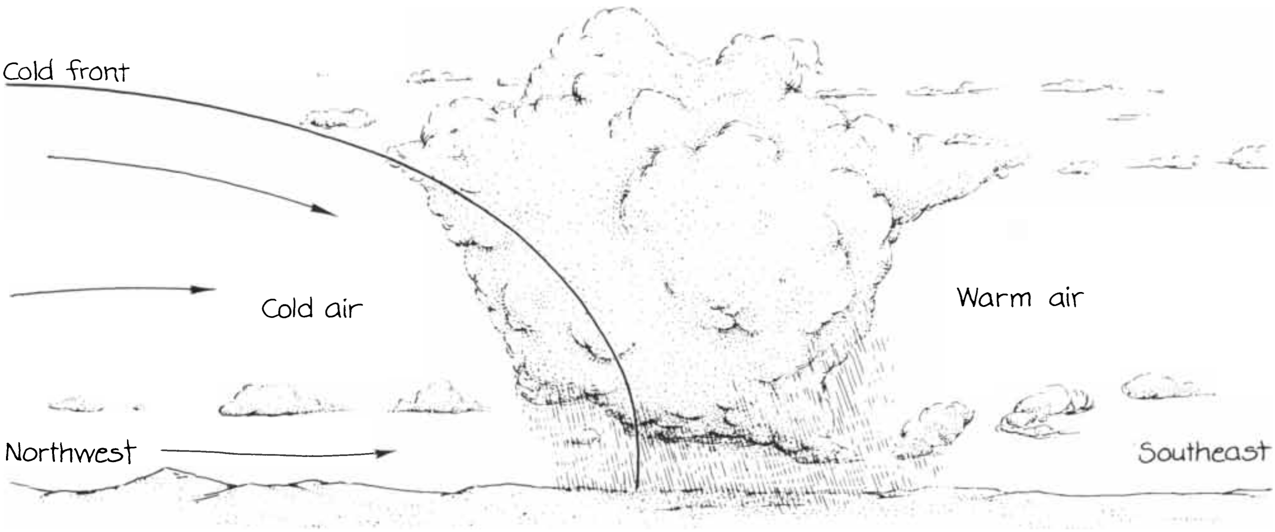
A coastal cyclone, the fourth general case, is peculiar to the coastal areas of the northeastern U.S. Its low-pressure center passes to the south and east of the station. Thus no clearly defined weather front is seen. Sometimes these storms are weak. Sometimes a coastal cyclone intensifies and becomes a "nor'easter," a type of storm that can last for days.

Stimets charted the average power spectrum for eight cold-front storms. The chart shows a large peak at 480 minutes (eight hours); it is the result of dividing the data into eight-hour periods and also reflects the fact that many of the storms lasted for about eight hours. The peak at 135 minutes is more important: it suggests a pattern in the rainfall rate. As the cold fronts passed over the recording station their rainfall rates tended to vary, producing a period of about 135 minutes. A smaller peak that appears at about 48 minutes may also represent a periodic variation.

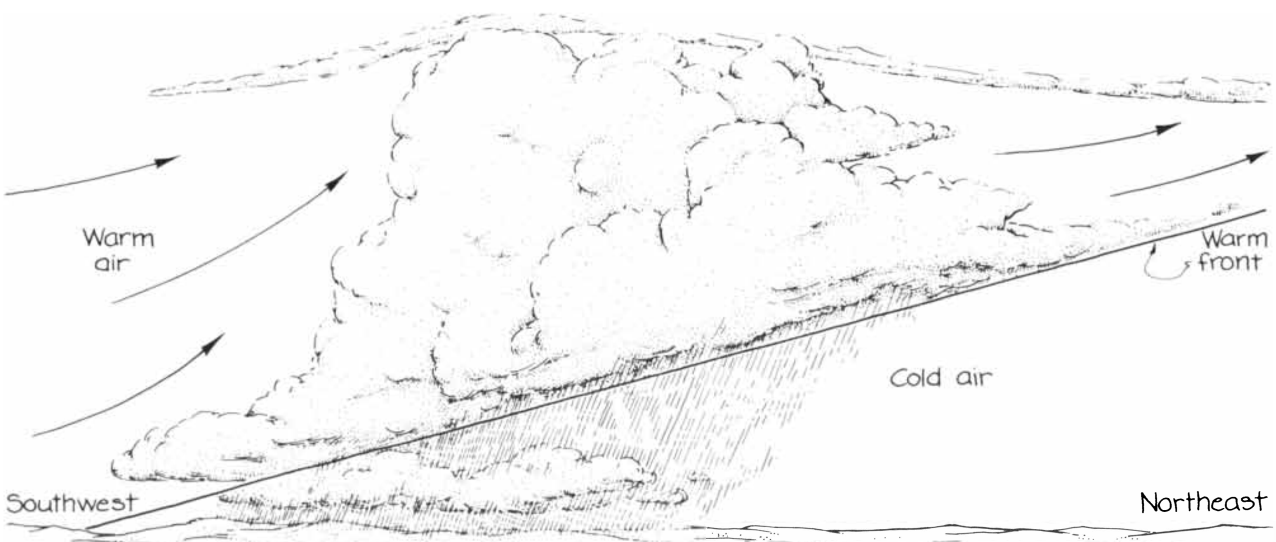
The power spectrum derived from the data collected during the passage of five warm fronts also reveals a characteristic periodicity. The power distribution again peaks at 480 minutes for reasons associated with the analysis of the data. More interesting are the peaks at 95 and 48 minutes. The data for six land-cyclone storms have a peak at 80 minutes. The data for 11 coastal storms appear to be entirely random.

Stimets notes that even though characteristic periods seem to be associated with storms in three of his four general categories, the organized component of the rainfall is minor. Most of the rainfall rate is random. The maximum organization appears to be associated with cold-front storms: up to 30 percent of their power lies in the long-period peaks, which rise out of the noise level.

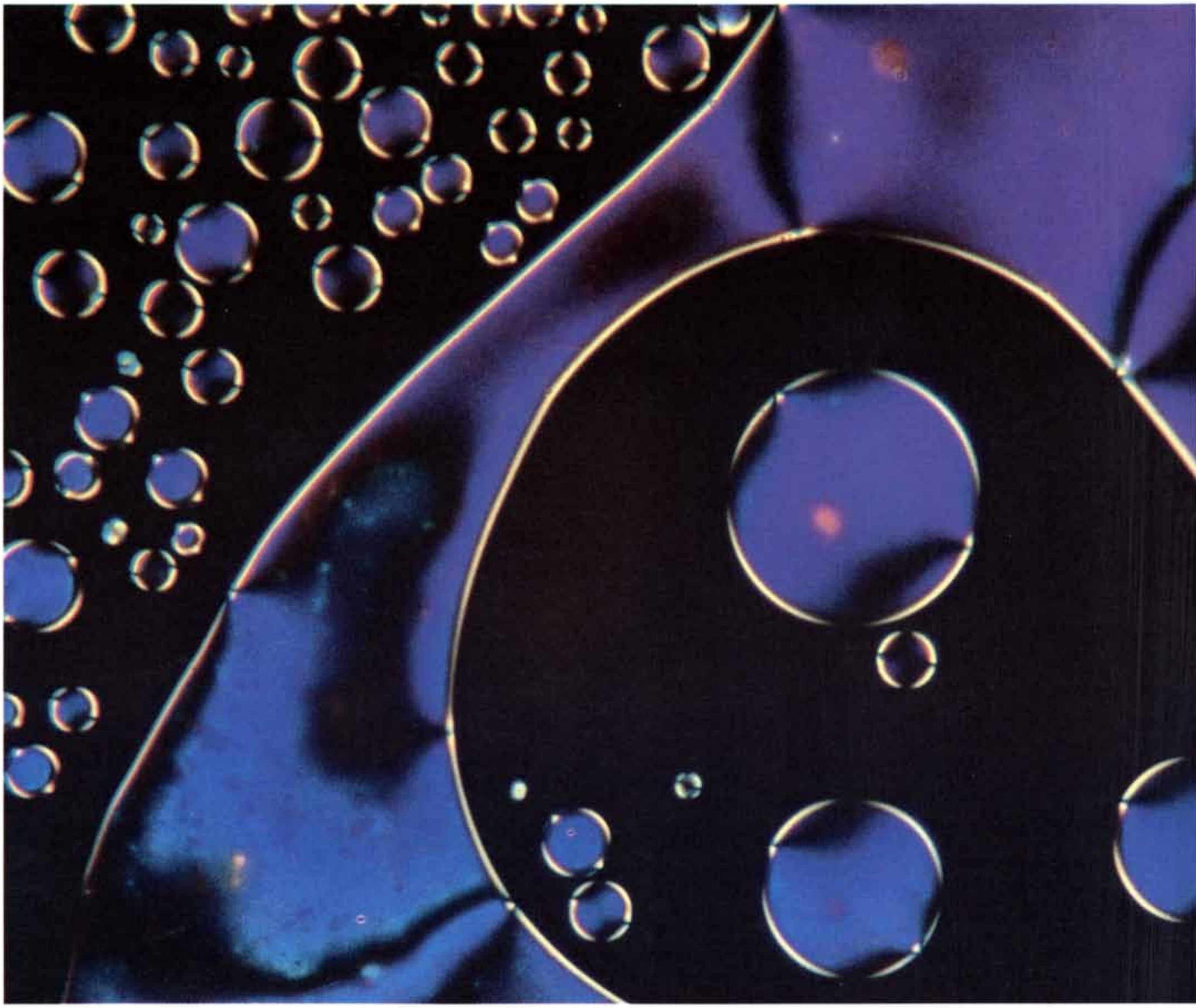
If indeed there is some organization to the rainfall in certain types of storms, does the spatial organization of the storm cause it? The question is accessible to observation. If the con-



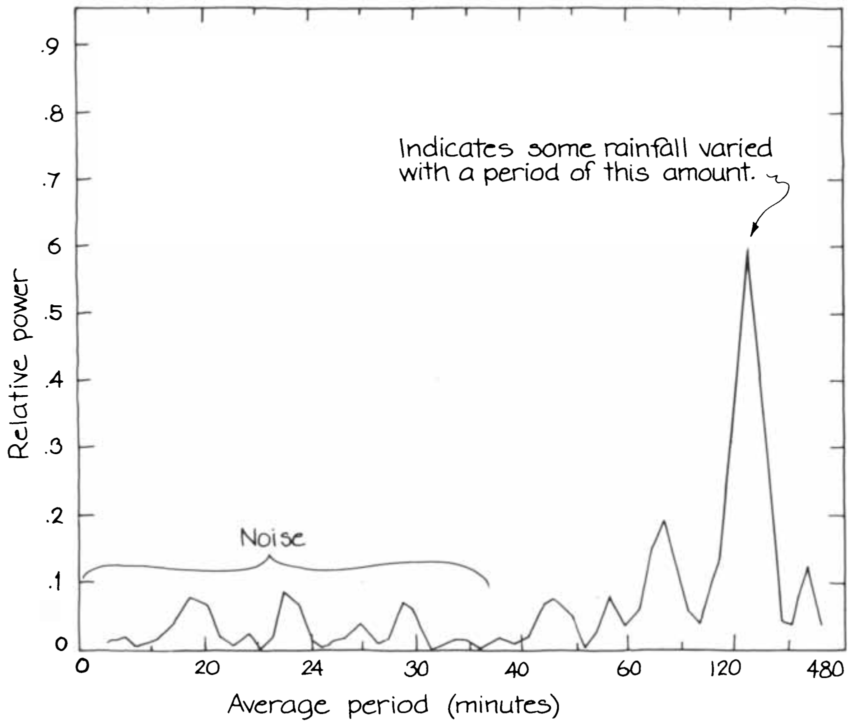

*A cross section through a cold front*



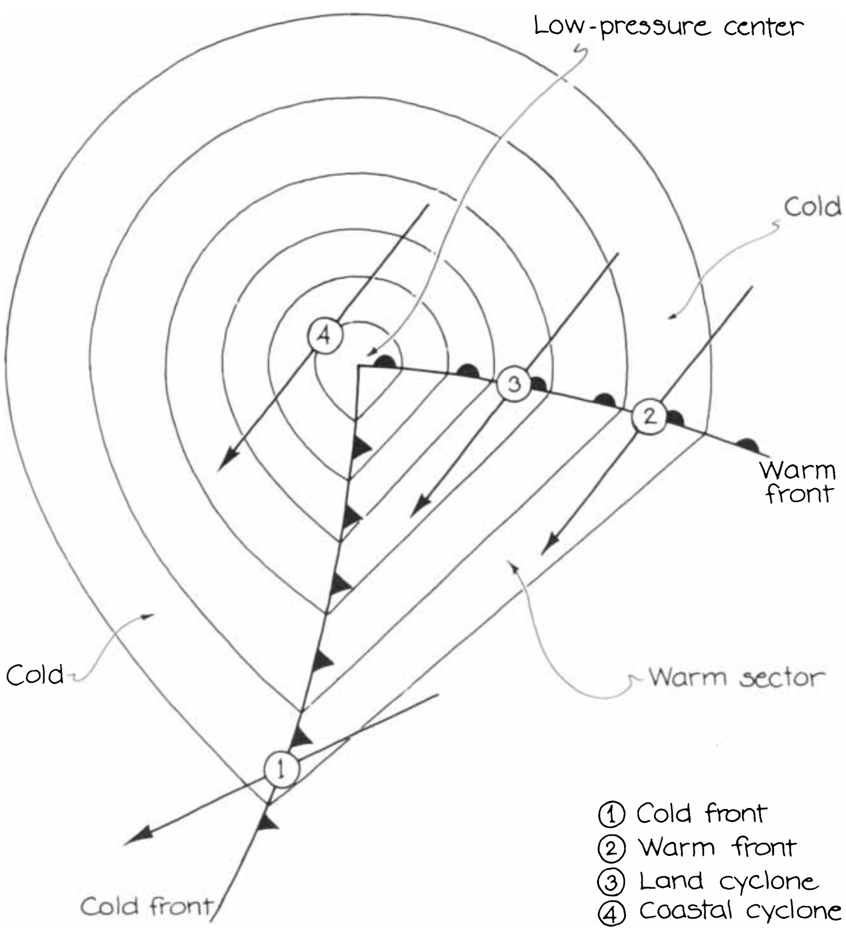

*A cross section through a warm front*



*This microphoto reveals the fluid nature of liquid crystals—a property Hitachi has put to colorful use in digital clocks, computer/television video screens, and a variety of innovative display devices.*




*The power spectrum of a storm*




*A recording station experiences four low-pressure systems*

centration of precipitation within the storm clouds does indeed vary in some periodic way, a ground station would record an intensification of rainfall as the formation responsible for it passed overhead.

Stimets points out that studies with radar and other instruments have indeed identified two basic precipitation-producing structures in storms: rain bands and precipitation cores. Rain bands are precipitation areas that may be hundreds of kilometers long. Distances of 50 to 100 kilometers separate rain bands from one another. They move in the approximate direction of the average airflow. Precipitation cores are smaller regions of concentrated precipitation that lie within rain bands.

The period between the passage of one rain band or core and the next is equal to the distance separating the structures divided by the speed at which they move. Stimets says the rain bands forming near cold fronts usually travel at speeds of 30 to 60 kilometers per hour. A spacing of 50 or 100 kilometers between rain bands would produce periods of between 50 and 200 minutes in the rainfall rates recorded at a station. The long-period peak at 135 minutes found in Stimets' data for cold-front storms falls quite neatly into this range. The smaller peak at 48 minutes is near enough to be considered also. He suggests that the long-period peak results from rain bands and that the smaller-period peak is more likely to be associated with the precipitation cores.

The average power spectrum for warm fronts has a peak at 95 minutes. For land cyclones the peak is at 80 minutes. Although both times fall within the range of possible periods for rain bands, Stimets believes they are more likely to be attributable to the precipitation cores.

The second peak in the data for warm fronts is at 48 minutes. It might indicate structure in the front passing over a recording station in about that much time. The peak might also be generated artificially in the analysis of the data. When the data contain narrow signals that are much stronger than the rest of the signals, a Fourier analysis can produce extra peaks that are harmonics of the basic peak. Since the main peak in the power spectrum for a warm front is at 95 minutes, the second harmonic would be expected at half that interval; the second harmonic almost matches the peak at 48 minutes.

Stimets attributes the absence of peaks in the coastal-storm data to the fact that no fronts passed over the stations. It is certainly conceivable that all the organization in a storm derives from a front. The variability of the coastal cyclones might also be responsible for the absence of periodicity. These storms seem more subject to changes in intensity and direction of travel than other storms. As a result any evidence of structure may be smeared.

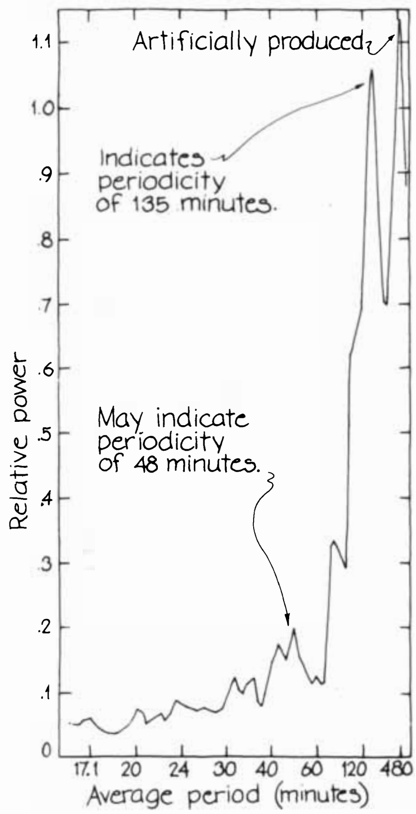
An improvement in instrumentation may lead to an improvement in the quality of information. Stimets has therefore recently begun to replace the strip-chart recorder in his equipment with a printing calculator to which the circuit shown in the illustration on the next page has been added. The circuit, which was designed by Cesare C. DeLizza of the University of Lowell, consists of a crystal oscillator and two dividers that produce one pulse per second. Stimets enters an initial count into the calculator and closes the switch labeled S1. Thereafter the oscillator adds a count every second to the calculator.

A dual flip-flop circuit monitors the rain gauge. Whenever the gauge causes a short circuit by tipping, the circuit prints the time displayed on the calculator. The device has reserve battery power so that it can function even when a storm or a careless person turns off the electrical power.

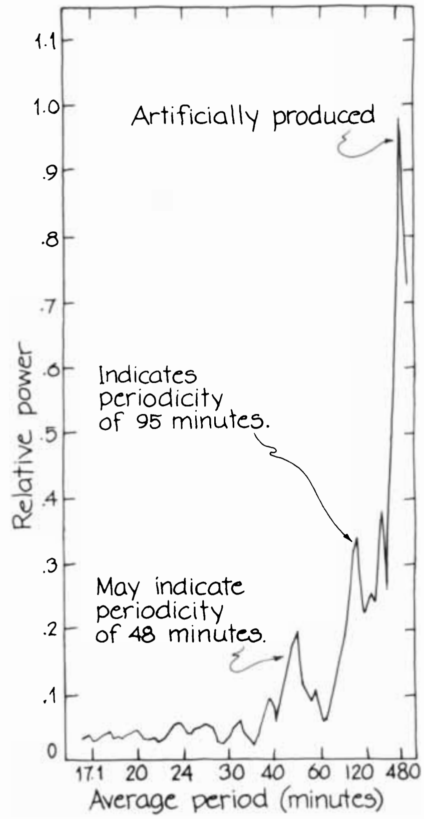
DeLizza's circuit is designed for a Sharp printing calculator. To attach it open the calculator and find where the leads from the "print" and the "+=" keys join the printed circuit board. Also find the common wire between those keys. Solder the "print" lead of DeLizza's circuit to the first of the junctions on the board, the "add 1" lead to the second junction and the "common" lead to the common wire between the keys. The connection represented by the top line in the circuit diagram is to be attached to the positive side (the red wire) of the calculator's adaptor socket for alternating current.

The connection represented by the bottom line in the diagram is to be attached to the negative terminal of the calculator battery. Solder a 1N4001 diode between the red and white wires in the adaptor socket for alternating current, so that the cathode connects with the red wire. The elements labeled TIL-111 are optically coupled phototransistors.

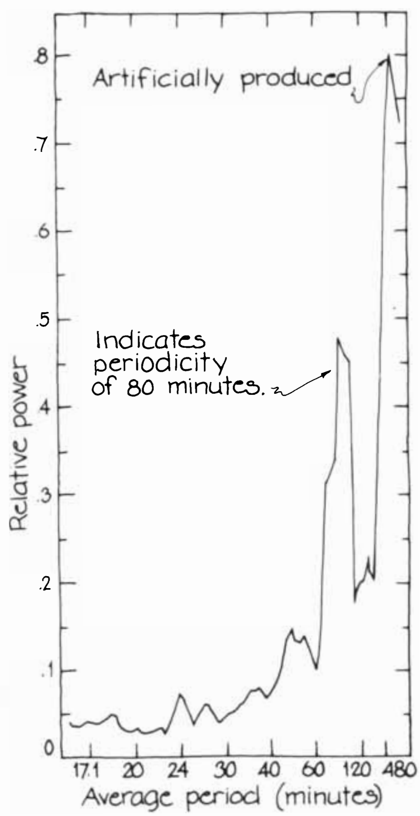
The element labeled X1 is a crystal that resonates at 3.579545 megahertz. Adjust the capacitor labeled C1 until a measurement of the frequency (made at pin 7 on the MM5369 clock oscillator) shows 3.579545 megahertz. The output of the clock oscillator is



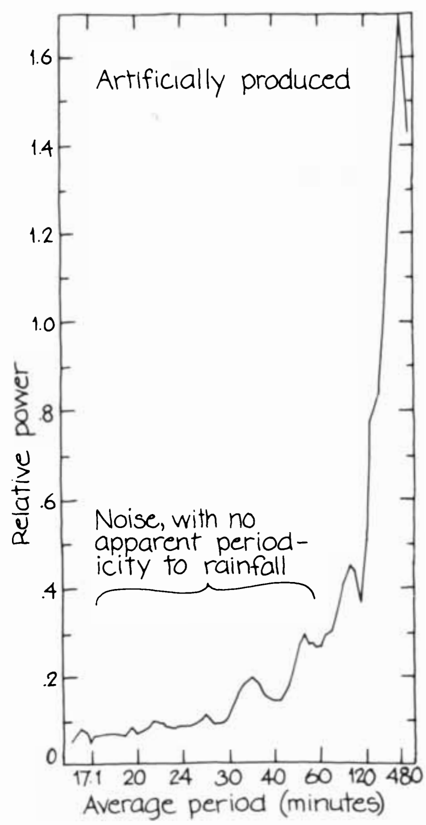

*Spectrum of cold-front storms*




*Spectrum of warm-front storms*




*Spectrum of land cyclones*




*Spectrum of coastal cyclones*

60 hertz. The first 4017 integrated circuit beyond it divides the frequency by 10. The next one divides it by 6. Thus the pulse rate communicated to the rest of the circuit is one hertz (one pulse per second).

The printing calculator gives a better resolution of variations in rainfall than the strip-chart recorder. Indeed, Stimets can now monitor individual clouds that take from five to seven minutes to pass. Preliminary results on the power spectrum of cold-front storms suggest that such a cloud may show up as a small peak.

Stimets is also working on installing a network of rain gauges to include many schools in the Boston area so that storms and the rain bands associated with them can be monitored over a much greater area.

In the past decade several research groups have discovered periodic structures in the storms associated with both warm and cold fronts. Peter V. Hobbs and his associates at the University of Washington have studied the large- and small-scale organization of middle-latitude cyclones. They employ an elaborate observation network composed of radar systems, instrumented airplanes and balloons that they coordinate with a system of ground stations.

Hobbs and his group studied in detail a cold front that passed through the observation network on November 17, 1976, moving east southeast. The precipitation near the front was concentrated in rain bands that were parallel or approximately parallel to the front. The bands showed three different arrangements.

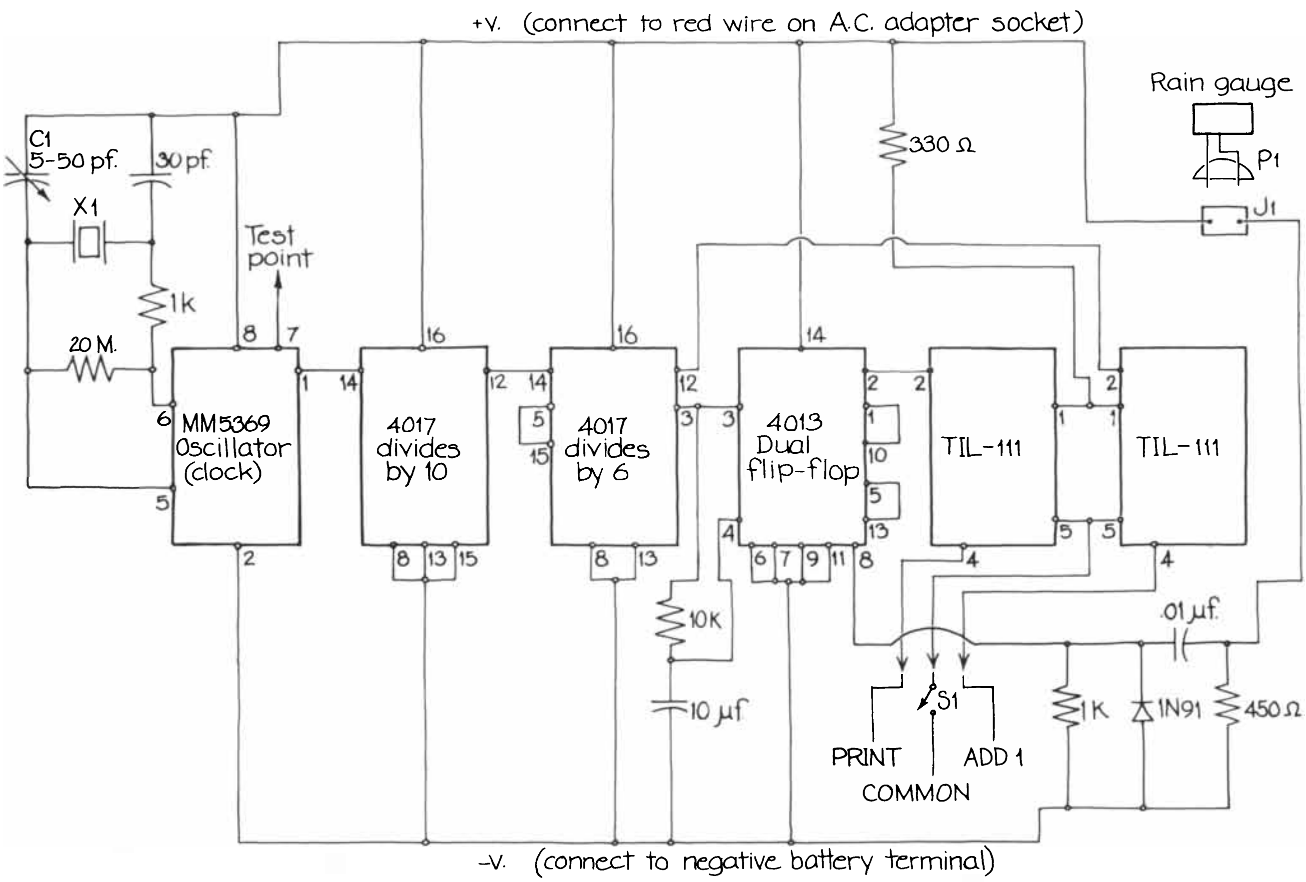
One arrangement extended over 50 kilometers of the warm sector ahead of the front. It consisted of two or three bands traveling faster than the cold front, the rear band being just ahead of the front at ground level. Another arrangement, behind the cold front, consisted of four bands that were each several tens of kilometers wide and traveled faster than the front. The third arrangement was a single rain band only four kilometers wide that stayed above the cold front.

The high points of this narrow band were at altitudes of from 1.5 to 4.5 kilometers. The other rain bands were higher: their peaks lay at altitudes of five and six kilometers. During the observation one of the wide bands behind the cold front overtook the front and passed it, apparently moving over the narrow rain band that travels with the front.

The rain bands in the warm sector and those behind the cold front produced concentrated precipitation from irregularly shaped cores. The narrow rain band along the front also had cores, but in that band they were elliptical and roughly aligned, suggesting that convection currents were more organized in these cores than they were elsewhere.

The rain bands and cores behind the cold front moved at roughly 100 kilometers per hour, identical with the wind speed that prevailed at altitudes of three to six kilometers. Evidently the wind steered the bands. The winds also steered the bands in the warm sector. The band above the ground location of the cold front moved at the speed of the front (about 65 kilometers per hour) rather than at the speed of the winds.

Air at the front of a rain band in the warm sector ascended rapidly and then flowed to the rear of the band. By tracing this movement, Hobbs and his group discerned a complicated system by which rain was produced. The top of the band functioned as a "seeder zone" of ice particles. Once the par-




*Circuitry for a printing calculator*

ticles nucleated they grew by the deposition of water vapor. Eventually they fell into a "feeder zone," where they grew further by sticking together, forming rime or receiving more water vapor. Below the feeder zone the temperature increased enough to melt the ice, producing rain.

The vertical structure of a wide rain band behind the cold front was similar except that it lacked the large-scale convection displayed by the warm sector preceding the front. As before, the particles initiated in the seeder zone grew and fell until they melted and became rain.

The seeder zones of both sets of rain bands contained regions of updrafts and downdrafts called generating cells. Ice particles apparently grew in the updraft portions until they were large enough to fall out of the cell; they may also have traveled laterally.

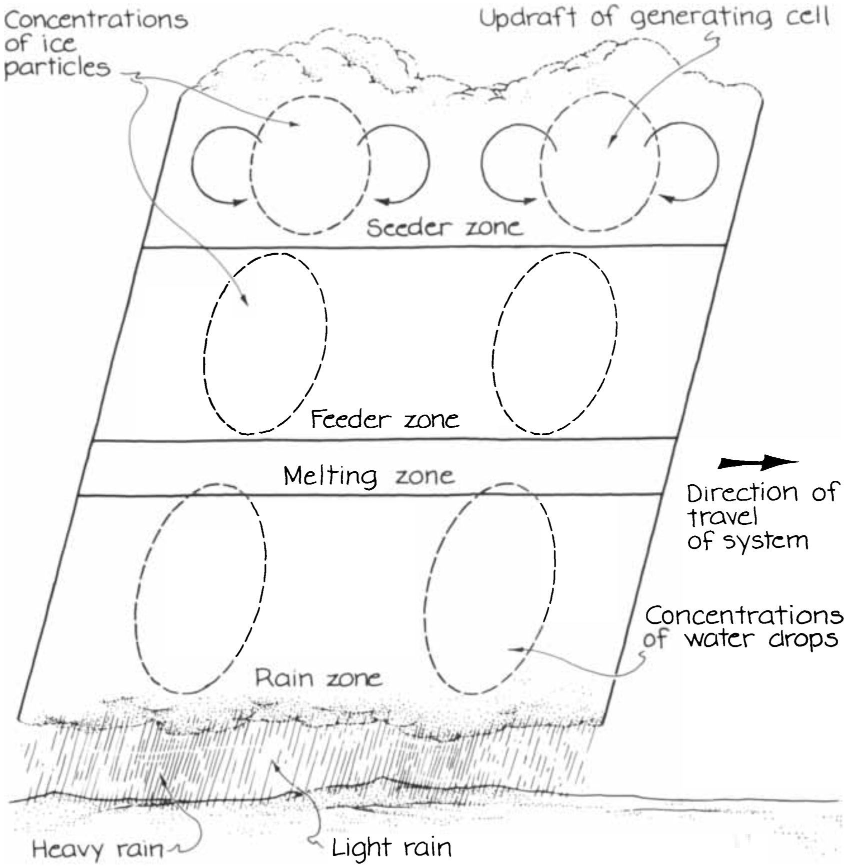
Apparently the horizontal arrangement of these generating cells parallel to the front produces the organization that leads to the formation of rain bands and presumably to the variations in rainfall detected by Stimets. Along a perpendicular line the cells align in a roughly periodic fashion. Hence rain bands and the variation in rainfall rates may be the result of the arrangement of generating cells in the seeder zones.

Can you detect rain bands without instruments? Imagine the variation in the rainfall when a cold-front storm of the kind described by Hobbs and his colleagues passes you. The first rain you detect is in the warm sector ahead of the front. The rainfall rate increases whenever a rain band in that front group passes you.

When the cold front reaches you, the rain band traveling with it delivers another heavy deluge. Whenever one of the wide rain bands behind the front passes, you again receive heavy rain. Eventually the last rain band of the storm system passes and the rainfall begins to let up.

Actually the rate of rainfall would not be that clearly organized. Superimposed on the general pattern of rainfall are smaller-scale variations due to the precipitation cores formed by the generating cells. The rainfall is also modified by winds through which the particles fall. At ground level in any given storm you might not always notice any general organization of the precipitation.

Can the results produced by Stimets be explained in terms of the cold-front storm investigated by Hobbs and his associates? Since Stimets worked with storms on the opposite side of America, such a comparison may not be fair.



*An idealized arrangement of rain bands*

Still, I wondered if the rainfall studied by Hobbs showed a characteristic variation over time.

I considered the rain bands traveling behind the cold front. Suppose they were uniformly spaced with a period of 30 kilometers. Traveling at about 100 kilometers per hour, they would yield a periodic variation in rainfall with a characteristic time of about 33 minutes.

The precipitation cores within the bands were more closely spaced than the bands. Since the cores were irregularly shaped, assigning them a characteristic spacing is difficult. Suppose it was about 10 kilometers. In that case the cores would yield variations in rainfall with a characteristic time of about six minutes.

Both my results suggest faster variations than one would infer from Stimets' data. Cold-front storms are highly variable, however, and my comparison may not be valid. Perhaps the rain bands studied by Hobbs and the other investigators moved faster than the ones studied by Stimets. One result of the higher speed would be to force greater variations in the rainfall.

Hobbs and his former student Paul H. Herzegh have reported much smaller structures, resembling rain bands, that are near a warm front. Radar observations revealed wavelike bands that formed near a stationary warm front running east and west. They were parallel to the front, about eight kilometers wide and spaced about 12 kilometers apart.

Hobbs and Herzegh found that the bands originated from generating cells in a seeder zone. The cells were laid out along a line perpendicular to the front. The updraft portions of the cells favored the formation of ice particles. Thus the roughly periodic arrangement of generating cells dictated the formation of the rain bands.

You might like to construct a rain gauge similar to Stimets'. Do the storm systems passing through your area display rain bands that can be recorded by your gauge? Do they display characteristic times of the kind detected by Stimets? A network of gauges deployed over a few hundred kilometers might enable you to track the rain bands. Both Stimets and I would be interested in your results. Questions about details of Stimets' work can be addressed to him at the Department of Physics, University of Lowell, Lowell, Mass. 01854.

# BIBLIOGRAPHY

*Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.*


## COMPUTER RECREATIONS

UNDERSTANDING NATURAL LANGUAGE. Terry Winograd. Academic Press, 1972.

EXPERIMENTS IN ARTIFICIAL INTELLIGENCE FOR SMALL COMPUTERS. John Krutch. Howard W. Sams & Co., Inc., 1982.

THE POLICEMAN'S BEARD IS HALF-CONSTRUCTED. Racter, with an introduction by William Chamberlain. Warner Books, Inc., 1984.

## THE COMMAND AND CONTROL OF NUCLEAR WAR

CAN NUCLEAR WAR BE CONTROLLED? Desmond Ball. International Institute for Strategic Studies, London, 1981.

THE COMMAND AND CONTROL OF NUCLEAR FORCES. Paul Bracken. Yale University Press, 1983.

STRATEGIC COMMAND, CONTROL, COMMUNICATIONS, AND INTELLIGENCE. Charles A. Zraket in *Science,* Vol. 224, No. 4655, pages 1306–1311; June 22, 1984.

## COLLAPSE AND FORMATION OF STARS

PROTOSTARS AND PLANETS: STUDIES OF STAR FORMATION AND OF THE ORIGIN OF THE SOLAR SYSTEM. Edited by Tom Gehrels. University of Arizona Press, 1978.

HYDRODYNAMIC COLLAPSE. Joel E. Tohline in *Fundamentals of Cosmic Physics,* Vol. 8, No. 1, pages 1–82; December, 1982.

## THE PERCEPTION OF SPEECH IN EARLY INFANCY

INFANT SPEECH PERCEPTION: A CRITICAL APPRAISAL. P. W. Jusczyk in *Perspectives of the Study of Speech,* edited by Peter D. Eimas and Joanne L. Miller. Lawrence Erlbaum Associates, Inc., 1981.

ON FINDING THAT SPEECH IS SPECIAL. Alvin M. Liberman in *American Psychologist,* Vol. 37, No. 2, pages 148–167; February, 1982.

## ALZHEIMER'S DISEASE

ALZHEIMER'S DISEASE: A REPORT OF PROGRESS IN RESEARCH. Edited by Suzanne Corkin, Kenneth Davis, John Growdon, Earl Usdin and Richard Wurtman. Raven Press, 1982.

BIOLOGICAL ASPECTS OF ALZHEIMER'S DISEASE (BANBURY REPORT 15). Edited by Robert Katzman. Cold Spring Harbor Laboratory, 1983.

CHOLINE AND CHOLINERGIC NEURONS. Jan K. Blusztajn and Richard J. Wurtman in *Science,* Vol. 221, No. 4611, pages 614–620; August 12, 1983.

ALZHEIMER'S DISEASE: ADVANCES IN BASIC RESEARCH AND THERAPIES. Edited by Richard J. Wurtman, Suzanne H. Corkin and John H. Growdon. Center for Brain Sciences and Metabolism Charitable Trust, P.O. Box 64, Cambridge, Mass., 1984.

## HOT NUCLEAR MATTER

CLASSICAL MICROSCOPIC CALCULATIONS OF HIGH-ENERGY COLLISIONS OF HEAVY IONS. A. R. Bodmer and C. N. Panos in *Physical Review C,* Vol. 15, No. 4, pages 1342–1358; April, 1977.

COLLECTIVE SIDEWARD FLOW OF NUCLEAR MATTER IN VIOLENT HIGH-ENERGY HEAVY-ION COLLISIONS. Horst Stöcker, Joachim A. Maruhn and Walter Greiner in *Physical Review Letters,* Vol. 44, No. 11, pages 725–728; March 17, 1980.

DENSITY PATTERNS AND ENERGY-ANGLE DISTRIBUTIONS FROM A SIMPLE CASCADE SCHEME FOR FAST $^{20}$NE + $^{238}$U COLLISIONS. E. C. Halbert in *Physical Review C,* Vol. 23, No. 1, pages 295–330; January, 1981.

COMPRESSION EFFECTS IN RELATIVISTIC NUCLEUS-NUCLEUS COLLISIONS. R. Stone, R. Bock, R. Brockman, J. W. Harris, A. Sandoval, H. Stroebel, K. L. Wolf, H. G. Pugh, L. S. Schroeder, M. Maier, R. E. Renfordt, A. Dacal and M. E. Ortiz in *Physical Review Letters,* Vol. 49, No. 17, pages 1236–1239; October 25, 1982.

## THE RELIABILITY OF COMPUTER MEMORIES

THE THEORY OF INFORMATION AND CODING. Robert J. McEliece. Addison-Wesley Publishing Co., Inc., 1978.

SOFT ERRORS IN VLSI: PRESENT AND FUTURE. Timothy C. May in *IEEE Transactions on Components, Hybrids, and Manufacturing Technology,* Vol. CHMT-2, No. 4, pages 377–387; December, 1979.

ERROR-CORRECTING CODES FOR SEMICONDUCTOR MEMORY APPLICATIONS: A STATE-OF-THE-ART REVIEW. C. L. Chen and M. Y. Hsiao in *IBM Journal of Research and Development,* Vol. 28, No. 2, pages 124–134; March, 1984.

## JET-PROPELLED SWIMMING IN SQUIDS

MECHANICAL DESIGN IN ORGANISMS. S. A. Wainwright, W. D. Biggs, J. D. Currey and J. M. Gosline. Edward Arnold, 1976.

MECHANICS AND ENERGETICS OF ANIMAL LOCOMOTION. Edited by R. McN. Alexander and G. Goldspink. Chapman & Hall, Ltd., 1977.

SQUID MANTLE MUSCLE. Q. Bone, A. Pulsford and A. D. Chubb in *Journal of the Marine Biological Association of the U.K.,* Vol. 61, No. 2, pages 327–342; May 1, 1981.

RESPIRATORY METABOLISM AND SWIMMING PERFORMANCE OF THE SQUID, *LOLIGO OPALESCENS.* R. K. O'Dor in *Canadian Journal of Fisheries and Aquatic Sciences,* Vol. 39, No. 4, pages 580–587; April, 1982.

THE ROLE OF ELASTIC ENERGY STORAGE MECHANISMS IN SWIMMING: AN ANALYSIS OF MANTLE ELASTICITY IN ESCAPE JETTING IN THE SQUID, *LOLIGO OPALESCENS.* John M. Gosline and Robert E. Shadwick in *Canadian Journal of Zoology,* Vol. 61, No. 6, pages 1421–1431; June, 1983.

## THE CROSSBOW

THE CROSSBOW, MEDIAEVAL AND MODERN, MILITARY AND SPORTING: ITS CONSTRUCTION, HISTORY AND MANAGEMENT. Ralph Payne-Gallwey. Bramhall House, New York, 1958.

CROSSBOWS: "FROM THIRTY-FIVE YEARS WITH THE WEAPON." George M. Stevens. Crossbow Books, Huntsville, Ark., 1978.

## THE AMATEUR SCIENTIST

THE MESOSCALE AND MICROSCALE STRUCTURE AND ORGANIZATION OF CLOUDS AND PRECIPITATION IN MIDLATITUDE CYCLONES, 1: A CASE STUDY OF A COLD FRONT. Peter V. Hobbs, Thomas J. Matejka, Paul H. Herzegh, John D. Locatelli and Robert A. Houze, Jr., in *Journal of the Atmospheric Sciences,* Vol. 37, No. 3, pages 568–596; March, 1980.

THE MESOSCALE AND MICROSCALE STRUCTURE AND ORGANIZATION OF CLOUDS AND PRECIPITATION IN MIDLATITUDE CYCLONES, 2: WARM-FRONTAL CLOUDS. Paul H. Herzegh and Peter V. Hobbs in *Journal of the Atmospheric Sciences,* Vol. 37, No. 3, pages 597–611; March, 1980.