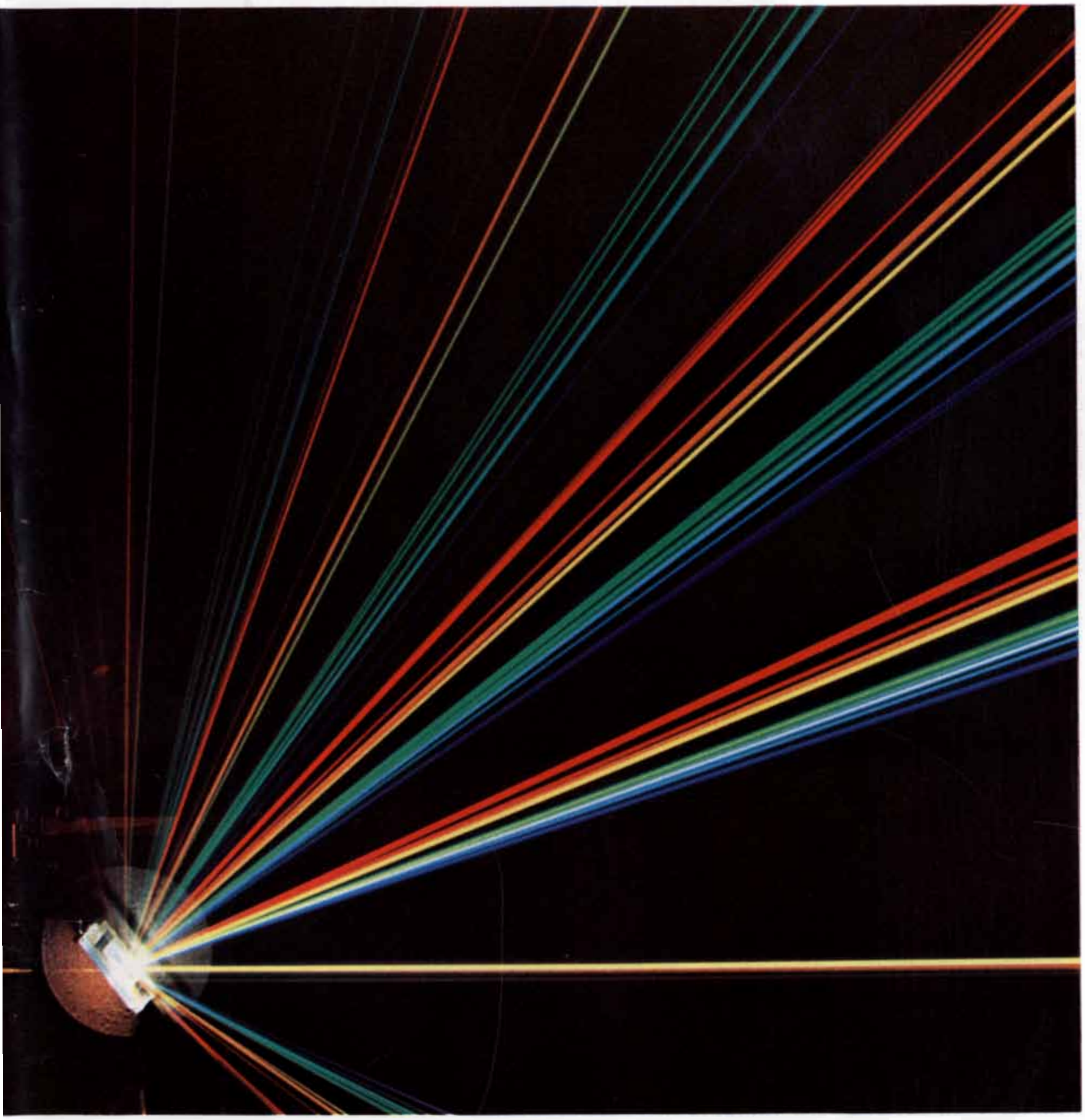
SCIENTIFIC AMERICAN



METAL-VAPOR LASERS

ONE DOLLAR

February 1973

The new Camaro Type LT. You deserve a car this good.

Camaro Type LT at Zuma Beach, California.

For 1973, there's a new kind of Camaro called the Type LT. The "L" is for luxury, the "T" is for touring.

In the realm of luxury, LT standards include full-foam, cloth-and-vinyl seats, power steering, wood-grain vinyl accents on the instrument panel and doors and Hide-A-Way windshield wipers.

For an even quieter, more luxurious ride, insulation has been added from the road to the roof.

And for the touring side of LT, there's Camaro's already-famous road-hugging suspension, Rally Wheels, twin sport mirrors (left-side is remote control), a Turbo-Fire 350 V8 and special instrumentation.

New Camaro Type LT. Now you can have your luxury and drive it, too.

1973 Chevrolet. Building a better way to see the U.S.A. Chevrolet

The Electronic Scratch-Pad for scientists and engineers. $495.

The technical problems you handle are tough enough without unnecessary distractions and delays. Like the inaccuracy of a slide rule (!) Or the time wasted with a mechanical calculator. And the bothersome chore of looking up log and trig functions in various tables.

So Victor designed a calculator that helps you keep your mind on the overall problem by taking care of the complex mathematical operations for you. Eleven digits more accurately than a slide rule. Considerably faster than a mechanical calculator. With instant, push-button access to log and trig functions accurate to 14 digits. Without interpolation.

What's more, it's even capable of handling long sequential operations exactly as they're written. So you don't have to break a problem down and store intermediate answers.

All this without sacrificing the practical features of a 4-function calculator or the rugged dependability of a compact, desk-top model. And all for $495.

Victor's Electronic Scratch-Pad. It takes care of the figures while you take care of the problems.

For more complete information on Victor's Series 1800 Scientific/ Engineering Electronic Calculator, simply call the Victor office listed in your Yellow Pages or send us the coupon.

VICTOR ELECTRONIC CALCULATORS

Victor Comptometer Corporation
Business Machines Division
3900 North Rockwell Street,
Chicago, Illinois 60618

Please send me full details about your New Series 1800 Scientific/ Engineering Electronic Calculator.

Name ____________________

Title ____________________

Firm ____________________

Address ____________________

City __________ State __________ Zip __________

Victor Comptometer Corporation, Chicago, Ill. 60618
Also Galt, Ontario, Canada and Zurich, Switzerland.
SA-23

SCIENCE/SCOPE

An AWG-9 Phoenix weapon control system, normally used for launching missiles from the Navy's F-14 Tomcat fighter, has been installed and tested in a shipboard defense role aboard the USNS Wheeling, where it successfully detected and tracked multiple targets at both high and low altitudes from the ship's deck. In multiple target tests, five aircraft were flown in the target area and successfully tracked. The Hughes-built AWG-9 system is unique in its ability to acquire and track more than 20 targets at the same time, launch up to six Phoenix missiles, and guide them simultaneously.

A new electronics fabrication technique sandwiches a very thin dielectric supporting an etched stripline center conductor between two thin air-filled sheet metal ground planes stamped in a configuration that assures optimal support and bonding to the dielectric sheet and suppresses undesired parallel plate radiation modes. Used for the corporate feed of large antenna systems, air-filled strip transmission line has proved superior to dielectric-filled stripline in experiments recently completed by Hughes engineers. Air-filled stripline has better electrical characteristics and is lighter in weight and considerably less expensive to produce.

The Army's Advanced Attack Helicopter (AAH) program has designated the Hughes-built TOW anti-tank missile as primary armament. Hughes is offering major helicopter prime contractors competing for the award complete fire-control system integration for both missile and gun, gunner's and pilot's night vision equipment, laser target designator, and total ground support.

Hughes' Electron Dynamics Division needs engineers for R&D programs to meet projected future demand for transistorized microwave amplifiers; Gunn and Impatt diode sources; microwave and millimeter wave mixers, detectors, and ferrite devices; traveling wave tubes for space applications; and high-voltage, high-efficiency microwave power supplies. Write: B. E. Shryack, Hughes Aircraft Company, 3100 W. Lomita Blvd., Torrance, CA 90509. Hughes is an equal opportunity M/F employer.

A reliable lightweight, low-cost radar for air-superiority fighter planes is being developed by Hughes in a multimillion-dollar company-funded program. Initially, the system is to be designed for Northrop's P-530. It will have a look-up, look-down, clutter-free display capability. Designed for air-to-air and air-to ground missions, it will provide the fire control function for the Cobra's guns, missiles, rockets, and bombs. Special emphasis is being placed on minimum maintenance.

The government of Iran has made its second major purchase of Hughes-built TOW anti-tank missiles from the U.S. Army Missile Command, Huntsville, Ala. Iran plans to deploy TOW with armored infantry, helicopter, and infantry units. The Netherlands, West Germany, and Italy have also chosen TOW, and several other countries are evaluating it for both ground and helicopter applications.

Creating a new world with electronics

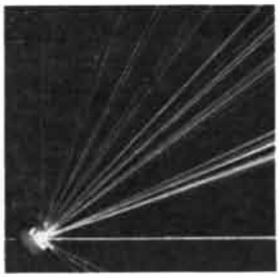
HUGHES

HUGHES AIRCRAFT COMPANY

Established 1845 **SCIENTIFIC AMERICAN** February 1973 Volume 228 Number 2

ARTICLES

DEPARTMENTS



THE COVER

The photograph on the cover shows the light produced by a helium-selenium laser, one of a class of lasers that operate with a vaporized metal mixed with another gas (see "Metal-Vapor Lasers," by William T. Silfvast, page 88). The output beam of the laser enters the picture at the lower left. As the beam passes through a small diffraction grating part of it is dispersed into its constituent colors, which are diffracted at various angles. The rest of the beam continues horizontally across the picture undispersed. The diffracted portion is separated into four distinct orders, or sets of colors, above the main beam and one order below the main beam. The first order fans out directly above the main beam, starting with blue and proceeding through the spectrum to red. The beams in the second, third and fourth orders, which continue over to the left side of the photograph, are progressively fainter but are progressively more widely dispersed. Thus they better display the large number of wavelengths generated by this single laser.

THE ILLUSTRATIONS

Cover photograph by F. W. Goro

Page	Source
15	U.S. Air Force
16–19	Dan Todd
20	U.S. Air Force
21	Texas Instruments Incorporated
22	National Aeronautics and Space Administration
24–25	Itek Corporation
27	Milford Small
28–29	Tom Prentiss
30	Marija Duda
31–35	Tom Prentiss
37	Francis G. Carey
38–40	Bunji Tagawa
41	Robin Ingle
42	Bunji Tagawa
43–44	Robin Ingle
50	Avery Brundage Collection, Center of Asian Art and Culture
52	Robin Ingle
53	Bernhard Karlgren and Jan Wirgin; characters in text by F. Y. Wang
54	Characters in text by F. Y. Wang
55	Characters in illustration by F. Y. Wang
56	F. Y. Wang and Robin Ingle
57	Robin Ingle
58	Robin Ingle from William S-Y. Wang
59	Robin Ingle from F. Y. Wang
64	John D. Woods
66	Allen Beechel
67–69	Graphic Presentation Services, Inc.
70	Allen Beechel
71	Graphic Presentation Services, Inc.
72	Allen Beechel
73	John D. Woods
74	Allen Beechel
75	Graphic Presentation Services, Inc.
79	Calspan Corporation
80–81	Dan Todd
82	Ben Rose (*top*), Calspan Corporation (*bottom*)
83	Ben Rose
84	Calspan Corporation
86	Dan Todd
88	F. W. Goro
90	Gabor Kiss
91	Gabor Kiss (*top*), F. W. Goro (*bottom*)
92–96	Gabor Kiss
99	Hale Observatories
100	Jim Egleson
101	Hale Observatories (*top*), Jim Egleson (*bottom*)
102–104	Jim Egleson
106–109	Ilil Arbel
110–115	Roger Hayward

LETTERS

Sirs:

Photoelastic analysis of the cross section of Gothic cathedrals could be a harmless exercise if the conclusions drawn by the author were not erroneous and irrelevant ["The Structural Analysis of Gothic Cathedrals," by Robert Mark; SCIENTIFIC AMERICAN, November, 1972]. This is so because the principal mode of behavior of the entire structure is dominated by rheological (instead of elastic) phenomena: long-term plastic flow and creep. Any serious investigation would also have to distinguish between the behavior of the structure during the relatively long period of construction, its initial, partial elastic response and the subsequent state affected by plastic flow, settlement and rotation of foundations. To assume that gravity and wind were suddenly "turned on" only after the completion of the cathedral is to believe in miracles in the wrong context. Needless to say, the physical characteristics of the materials used and the subsurface conditions must be known and taken into account. Photoelasticity is not an appropriate tool for such investigations.

Common sense coupled with a bit of physical intuition and supported by a few simple calculations might, however, justify the following conjectures: The builders of the cathedrals (apart from overriding aesthetic and spiritual considerations) were primarily preoccupied with gravity loads. I also suspect that they were much concerned with the construction process itself and were quite conscious of continual movements and differential settlements that occurred during and after construction. This might be the reason for the systematic differences in the width of mortar joints that are found in many Gothic cathedrals. This suggests that the flying buttresses might have been, at least in part, conceived as "scaffolding" to ensure stability during construction and in the early phase, but became (at least structurally) ineffective after settlement and plastic deformations had taken place.

This was probably understood by the builders, who were fortunately not misled by reading *Scientific American.* The aesthetic and scientific questions that arise are fascinating and complex, but the latter ones should yield to modern analytical techniques and thereby contribute to the better understanding of the intent of the cathedral builders.

PAUL WEIDLINGER

New York, N.Y.

Sirs:

Mr. Weidlinger has misinterpreted our use of analytical tools for architectural history and criticism. Indeed, over the period of seven centuries the Gothic buildings have undergone a very complex loading history, so that a complete description of their time-dependent response would require a more complex model. The problem is further compounded by a series of restorations wherein much of the critically loaded structure (for example stones at the ends of flying buttresses) has been replaced from time to time. Foundation movements have also contributed to failures.

However, the "relatively long period of construction" under discussion at Bourges and Chartres spanned only the years 1195 to 1221, and for the major portion of this period only vertical walls and piers were placed, so that lateral forces which might have caused rotation of foundations were almost nonexistent. Furthermore, as was often the situation, both cathedrals were built on the site of older, large buildings so that their foundations were more stable than if they had been placed on virgin soil. As in modern concrete construction, it hardly required a miracle to "turn on" gravity loadings from vaults and flying buttresses after these structures had been assembled on removable wooden scaffolding (see John Fitchen, *The Construction of Gothic Cathedrals,* Oxford University Press, London, 1961). By the same token, the wind loadings, which are largely derived from wind action on the last-constructed upper works, are also "turned on" following the removal of temporary construction scaffolding and with the arrival of major storms. It was these two effects that we modeled; the effect of the dead weight of the roof and its framing, piers and the longitudinal structure between the piers was treated analytically by considering static equilibrium. But the second part of Mr. Weidlinger's observation seems to be that if one is willing to wait for "settlement and plastic deformations [to have] taken place," then the equations of equilibrium no longer need apply. What miracle, then, will support the tall, thin piers against the lateral thrust of the vaults and wind loadings on the roof after the flying buttresses become structurally ineffective? As a matter of fact, as was noted by Jacques Heyman in a publication of the Newcomen Society (*Transactions,* XL, 1967–1968), the collapse of the choir at Beauvais Cathedral in 1284, following the construction period from 1247 to 1272, has been attributed to just such ineffectual buttressing.

The claim for photoelastic modeling is not that it gives *the* answer. Rather, that it provides a relatively simple engineering approach to what admittedly are very complex structures. Based on clearly stated assumptions, it allows us to locate critical regions (regions likely to exhibit tension) and to be able to make quantitative comparisons between the conceptual designs of two major buildings. I welcome suggestions of other modern analytical techniques that might help to further this understanding, but as for conjecture based on "common sense coupled with a bit of physical intuition and supported by a few simple calculations," there already exists in the extensive literature on the Gothic cathedral a glut of such opinions. It was the contradictions among these conjectures that led to our studies.

ROBERT MARK

Princeton University
Princeton, N.J.

Scientific American, February, 1973; Vol. 228, No. 2. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

Advertising correspondence should be addressed to Harry T. Morris, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 25 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

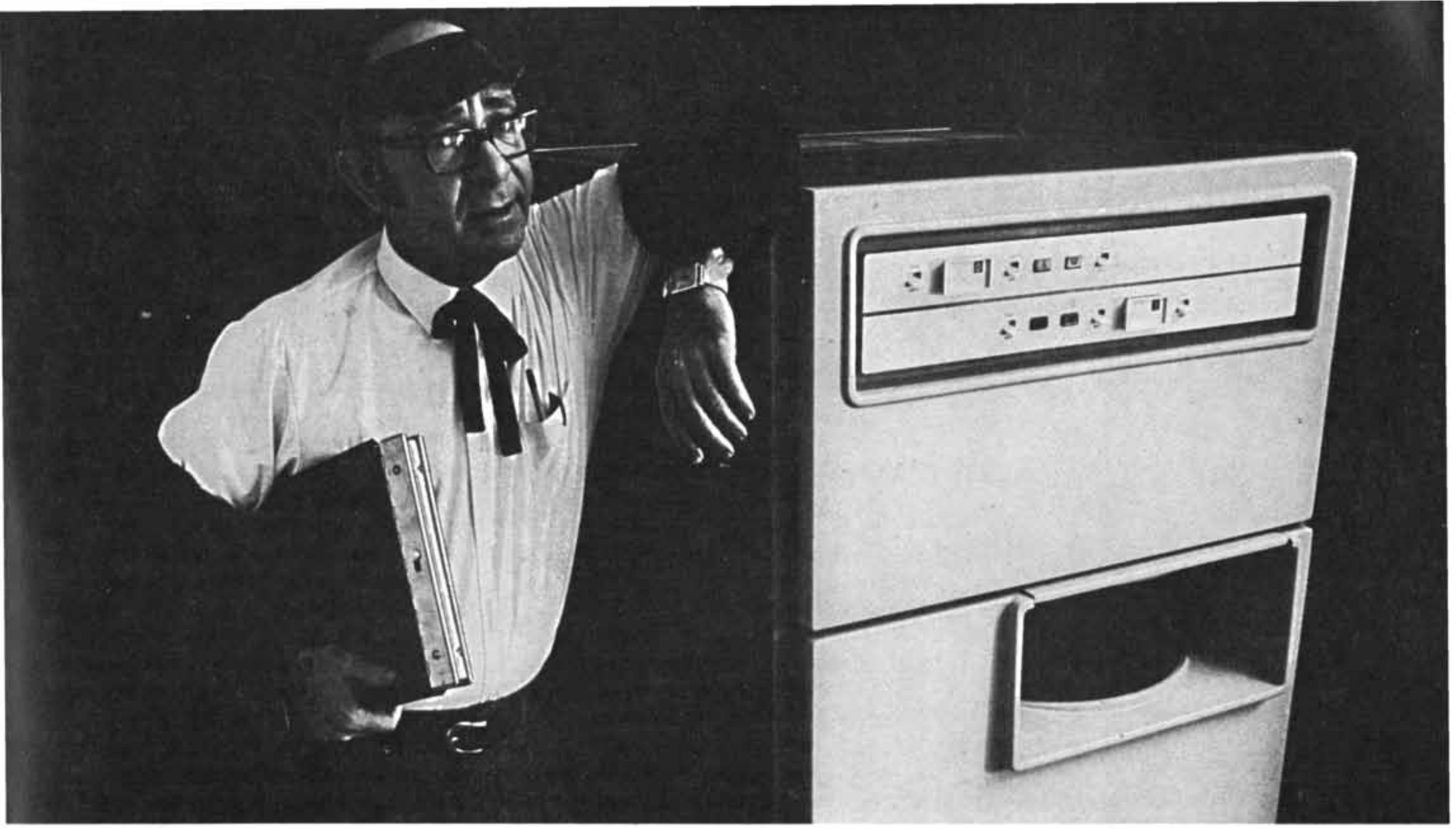
☐ Please enter change of address.

☐ Please enter subscription.
U.S. and Canada:
☐ 1 year, $10 ☐ 2 years, $18 ☐ 3 years, $25
All other countries:
☐ 1 year, $12.50 ☐ 2 years, $23 ☐ 3 years, $32

NAME

NEW ADDRESS

OLD ADDRESS



Buy your own disk drive.

Can you afford to? Can you afford not to?

We have a set of figures worked out by one of our accountants that shows you a savings of 32% over a three year period if you *buy* instead of lease.

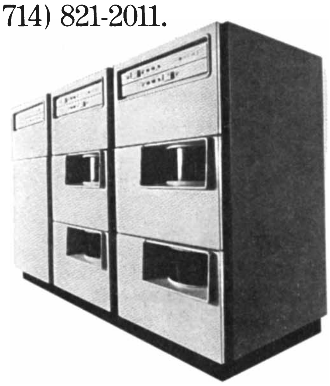
But your figures will be different from his. They always are.

Here's what you positively *can* afford to do. You can call (714) 821-2011 collect, and talk to Rod Murray. Get together with him or one of his people and your accountant.

Find out if you really can save something by buying instead of leasing.

But if you prefer to lease, check to see if you're leasing from us. If not, you're still spending too much.

Call your local CalComp office, or contact California Computer Products, Inc., SA-M2-73, 2411 West La Palma Avenue, Anaheim, California 92801. (714) 821-2011.



CALCOMP

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

FEBRUARY, 1923: "President Harding's Coal Commission as organized in Washington is beginning to get into action. Its functions and its limitations are to get the facts from the bottom to the top of the coal industry, both hard and soft, and report them to the President and to Congress. In April the greatest coal strike in the history of the world was staged. What the average man with a furnace or a coal stove has utterly failed to realize is that his home supply of coal is a very small portion of the coal problem. Our annual consumption of coal is about six tons for every individual, and less than one and a half tons of this are used for heating and cooking. We may quite overlook the large amount of coal that goes into each house over the electric wire, through the water pipes and as gas and ice. Even more is coal the staff of life of American industries. Every ton of steel produced has cost two tons of coal, every ton of cement has cost half a ton of coal—even a newspaper or magazine has cost more than twice its weight in coal."

"At present the Air Mail Service consists of a relay advance of mail from New York to San Francisco and vice versa. That is to say, the mail is picked up and carried ahead of the train mail, thereby gaining several hours at such cities as Cleveland and Chicago. The net result is that the Air Mail Service every day advances about 12,000 pounds of first-class letter mail by some three or four hours, and this, where the advance delivery is made late in the evening, means an actual advance of 15 to 18 hours. The post office authorities are realizing that to get the full service out of an airplane it should fly by night as well as by day. Experiments and study of the question have reached a stage where it is safe to conclude that it is entirely possible to fly at night. Within a few months a night journey will be made from Chicago to Cheyenne, Wyo. If that is successful, the carriage of mail will become a continuous movement, flying from New York to Chicago by day, from Chicago to Cheyenne by night and from Cheyenne to San Francisco during the early part of the second day. If the night experiment proves out, mail can be carried from New York to San Francisco in from 28 to 30 hours."

"The improvement in long-distance telephone communication in the past five years has been very noticeable. Today we in New York talk to Havana or San Francisco with almost no delay and perfect voice transmission, as part of the day's business. Vacuum-tube repeaters have made this possible. The repeaters are placed at regular intervals along the line, and as the currents become weakened they pick them up and deliver back into the line a current many times stronger. Other devices that provide for duplex use of the wires make them economical. The same wires that carry voices from city to city carry the news of the world to our daily newspapers, the stock reports to the centers of commerce and provide 'leased wire' telegraphic communication between business houses. The very recent application of the printing telegraph even further increases this wire capacity. With this device the message is received by a typewriter-like instrument instead of by an operator using code signals. Each pair of telephone wires can carry one telephone conversation, four ordinary telegraph circuits or eight printer telegraph circuits. Still another device has come into use to meet the demand for long-distance telephone lines. This is the carrier system that makes it possible to transmit over a single circuit a number of telephone or telegraph messages simultaneously."

SCIENTIFIC AMERICAN

FEBRUARY, 1873: "The question of heating our buildings by steam is now about in the same condition as was the question of lighting our dwellings by gas 50 years ago. It is true that gas is the safest kind of light, much safer than kerosene or oil or candles. All the same there were scores of people raising their voices against the introduction of gas in our cities and houses. We now have exactly the same kind of alarmists among us, who, notwithstanding that heating buildings by steam is the safest of all methods—far surpassing the hot-air furnaces and infinitely safer than having a separate fire in each room—raise their voices against steam heat and attribute every fire taking place in a locality where steam is used not to the fire in the furnace but to the steam, making the pipes red hot and so igniting the woodwork in the neighborhood of these pipes."

"Matthew Fontaine Maury, formerly an officer in the United States Navy, afterward of the Confederate Navy, died recently at his residence at Lexington, Va., aged 67. He was formerly superintendent of the Government Hydrographic Office, where he elaborated investigations in regard to winds and ocean currents. The discovery of the telegraphic ocean plateau and the indication of good whaling ground is attributed to him. At the time of his death he was professor of meteorology at the Virginia Military Institute."

"The presence of Miss Emily Faithfull in this country at the present time has revived the discussion of the question of women in society. A meeting was held a few evenings since at Steinway Hall in New York that must have given great encouragement to the advocates of the new movement. Miss Faithfull's address was reported in full in the newspapers and need not be repeated here, but some of the ideas suggested in it are deserving of careful study and consideration on the part of thoughtful citizens everywhere. One thing is certain: The right of woman to her share of honest labor cannot be put down by ridicule or despotism. It must be met fairly and squarely, and now that it has been taken up by our most refined and gifted women we trust that the question will soon be settled to the entire satisfaction of all parties."

"A recent heavy rain followed by a severe frost in New York City played great havoc with the telephone wires. The weight of ice was so great as to cause innumerable breaks, so that the elaborate fire alarm system became useless, and for a time it appeared as if communication with other cities was to be seriously interrupted. This has caused a re-agitation of the subject of underground wires, which, it is justly believed, would prove of great advantage as compared with those now in use. Not only would the unsightly telegraph poles be removed from our streets but also such casualties as the above, due to weather, would be permanently averted."

"A bill is now before Congress for the organization of a new territory, under the name of Oklahoma territory, out of lands now pertaining to the Indian territory."

Some things are changing for the better.

With the new HP-80, here's how easy it is to calculate the annual yield to maturity of a bond:

Purchase date: **March 15, 1973**	3 • 1 5 1 9 7 3	SAVE↑
Maturity date: **March 15, 2003**	3 • 1 5 2 0 0 3	DATE DAY
Coupon: **5%**	5	INTR PMT
Present value: **75**	7 5	BOND PV
Yield (YTM):		YTM i

With four data entries and six keystrokes, you get the answer: 7.01 (%)

The fundamental formula in common use by the financial community for the same problem follows. Solving for p (price of bond expressed as a percent) is tough enough; care to figure m (annual yield to maturity) without a computer?

$$p = -100\left(\frac{a}{2}\right)\left(\frac{180-b}{180}\right) + \left[\sum_{k=1}^{c} \frac{100\left(\frac{a}{2}\right)}{\left(1+\frac{m}{2}\right)^{\{k-1+(b/180)\}}}\right] + \frac{100}{\left(1+\frac{m}{2}\right)^{\{c-1+(b/180)\}}}$$

Now, a handful of fast solutions to time-and-money problems.

Most problems in business and finance are those involving time and money calculations: computing bond prices, rate of return on investments, depreciation, and the like.

Problems like these can take up to 20 minutes to solve, with constant referral to cumbersome tables. This means that there are numerous possibilities for error.

Now, time and money calculations can be done quickly, easily and accurately with the HP-80, the financial counterpart of the shirt-pocket-sized HP-35 calculator used by thousands of scientists and engineers.

The 9-ounce HP-80 has 40 specific capabilities built into its internal circuitry, which combine to perform more than 100 different financial calculations involving a relationship between time and money. There's no need to remember complicated formulas or use tables. All necessary programs, including a 200-year calendar, are built into the HP-80's solid-state memory.

You just enter the data, push a few keys, and read the answer – accurate to within the last penny on a million-dollar transaction.

Since it can perform most financial calculations in about 10 seconds, a user can save up to 20 minutes on a single problem. The price is $395 (U.S. only, plus tax).

If you're in business, you can't afford to be without it. Write for full information.

A better chance for the newborn.

Each year in the U.S. alone, almost a half-million babies begin life with a physical or mental impairment. Yet most medical authorities agree that a significant number of these tragedies can be prevented or mitigated, *right now,* by borrowing from hospital experience with patient monitoring techniques that were developed for intensive care units in cardiac, surgical and fetal care.

For the neonate the greatest need is to monitor respiration continuously. All too frequently, the newborn baby stops breathing spontaneously and unexpectedly – apnea, in medical terminology. A nurse usually can revive him just by picking him up or tickling his feet. But she must do so within 30 seconds or the infant runs a high risk of brain damage, even death.

Temperature and, in the high risk cases, blood pressure and ECG should also be monitored. Continuous monitoring not only helps prevent disaster in the nursery crib, but also aids the pediatrician in his diagnosis and therapy.

Building on our decade of experience with

Many people know us as an instrument manufacturer:
we make more than 2,000 products for measurement, test and analysis.
Others know us as a computer company: more than
10,000 own our calculators and computers. We prefer to think
that our business is to serve your measurement and computation needs.

cardiac, surgical and fetal monitoring systems, HP has recently introduced a compact new instrument designed expressly for neonatal monitoring. A plug-in module monitors respiration from a set of skin electrodes that detect changes in transthoracic impedance due to respiration; the same set of electrodes can be used to monitor ECG and heart rate. A second module monitors temperature to 0.1°. Both display the measurement in large red numbers that are easily read even in nurseries without a central station; they have adjustable alarms, visual and audible, to alert the nurse in case of trouble. The neonate monitor with respiration and temperature plug-ins is priced under $2000 and is completely compatible with HP's extensive ICU line. A complete description is available on request.

How to live happily in an increasingly digital world.

Most electronic products have changed drastically in the last few years. Everywhere you look, digital logic is displacing the familiar analog circuitry. And you'll be deceived if you believe that trouble-shooting these new products is the same as it was before the advent of the ubiquitous integrated circuit (IC). There are at least two important differences.

In the first place, you can easily check if any of the discrete components in an analog circuit is functioning properly by testing its input and output leads, usually two or, at worst, three. In a digital circuit, on the other hand, each IC has 14 or 16 leads and you may need to know what's going on at all the leads simultaneously.

Furthermore digital circuits are intentionally designed with low output impedances to make them insensitive to noise . . . perhaps their greatest single virtue. Unfortunately for circuit troubleshooters, this means that each IC input is clamped in either a high or a low logic state; hence the only way to perform a stimulus/response test on an IC has been to unsolder its input lead before injecting a test pulse and observing the output.

We're happy to report that things have changed for the better in the world of digital logic testing, with HP's IC Troubleshooters.

First there's the Logic Clip which instantaneously indicates the logic state of all 14 or 16 nodes on an IC without cables or power connections. Or the Logic Comparator which compares an in-circuit test IC with a duplicate IC known to be good and identifies any dissimilarities.

For testing individual IC nodes, there is a series of Logic Probes covering the various IC families which immediately indicate logic states and pulse activity. New partner to the Logic Probes is the Logic Pulser which automatically injects a single pulse that drives a logic-low node high and a logic-high node low for 300 nsec. These are an effective stimulus/response test team — without unsoldering leads and despite low IC output impedances.

We've told the full story in an easy-to-read booklet that's yours for the asking.

For more information on the products described in these pages, fill out the coupon or write to: Hewlett-Packard, 1505 Page Mill Road, Palo Alto, Calif. 94304.

HEWLETT hp PACKARD

Sales, service and support in 172 centers in 65 countries.

00303

Please send me information on the following:
( ) HP-80 Business Calculator
( ) HP Neonate Monitor
( ) IC Troubleshooters

Name____________________

Title____________________

Company____________________

Address____________________

City__________ State__________ Zip______

THE AUTHORS

TED GREENWOOD ("Reconnaissance and Arms Control") is completing work for his Ph.D. in political science at the Massachusetts Institute of Technology. His earlier degrees are in physics and mathematics: he obtained his bachelor's degree in mathematics and physics at the University of Toronto in 1967 and his master's degree in physics at M.I.T. in 1970. Noting an early interest in theology and philosophy as well as the natural sciences, he writes that "a resurgence of concern with political and policy issues several years ago led me to abandon physics." He adds that although he works primarily on arms control and national security, he is "also interested in many of the other policy areas where technology impinges on the public sector and requires a response or anticipation by government," an interest "that of course leads to an interest in the processes by which governments determine policies." Greenwood's article is based on a paper of his, "Reconnaissance, Surveillance and Arms Control," that was published last year by the International Institute for Strategic Studies.

MARCUS JACOBSON and R. KEVIN HUNT ("The Origins of Nerve-Cell Specificity") are respectively professor of biophysics at Johns Hopkins University and medical-scientist training program fellow in anatomy at the University of Pennsylvania. Jacobson, who was born in South Africa, studied science and medicine at the University of Cape Town and obtained his Ph.D. from the University of Edinburgh in 1960. He was a lecturer in physiology at Edinburgh for five years before he came to the U.S. His book *Developmental Neurobiology* was published in 1970. Hunt began collaborating with Jacobson at Johns Hopkins, where he received his bachelor's degree in biology and his master's degree in the humanities.

FRANCIS G. CAREY ("Fishes with Warm Bodies") is at the Woods Hole Oceanographic Institution. Describing himself as having been interested in biology since childhood, he writes: "I survived high school, attended William and Mary College and obtained a Ph.D. in biology from Harvard University in 1960. I have been at Woods Hole for some 10 years. After working at various projects of a biochemical nature, I learned that tunas are warm and have been fascinated by them ever since. I enjoy being on the water, and the work provides an opportunity for getting wet, pulling on ropes, gadgeteering, reading in the library and the aesthetic pleasure of working with these beautiful fishes."

WILLIAM S-Y. WANG ("The Chinese Language") is professor of linguistics at the University of California at Berkeley and director of its phonology laboratory. He is also director of the Project on Linguistic Analysis, which he founded in 1961 with support from the National Science Foundation. Wang received his early education in China, was graduated from Columbia College and obtained his Ph.D. from the University of Michigan. He and several colleagues have developed a computerized dictionary of Chinese dialects. In phonology his special interest is in the nature of tones; the work has taken him to many areas where tone systems are found, including Japan, Mexico, Yugoslavia, Southeast Asia and Scandinavia. He is editor of the newly established *Journal of Chinese Linguistics.* "On the lighter side," Wang writes, "I try to play the flute occasionally, limiting myself mostly to its middle register. During his more tolerant moments my son will do flute duets with me, which I enjoy especially much."

MICHAEL GREGG ("The Microstructure of the Ocean") is an assistant research oceanographer at the Scripps Institution of Oceanography. He also teaches a course in physical oceanography at San Diego State University. "My oceanographic career to date has been spent at Scripps," he writes, "as a graduate student from 1966 to 1971 and then as a postdoctoral associate until I received my present appointment in 1972. Before coming to Scripps I had a five-year tour in the Navy, all of which I spent at the division of naval reactors of the U.S. Atomic Energy Commission. My general interest is to develop instrumentation for critical measurements at sea." Gregg received his bachelor's degree (in physics) from Yale College in 1961.

PATRICK M. MILLER ("The Crashworthiness of Automobiles") is head of the structural dynamics section of the transportation research department of the Calspan Corporation, which was until recently the Cornell Aeronautical Laboratory, Inc. After receiving his bachelor's degree in mathematics from Michigan State University in 1957, he taught mathematics in high school for two years, returning to Michigan State for advanced degrees in applied mechanics: a master's degree in 1960 and a doctorate in 1966. His major activity for the past four years has been in a program of basic research in crashworthiness sponsored by the National Highway Traffic Safety Administration; much of the material in his article has been derived from that study. Miller writes that his outside interests "include model railroading, photography and restoring antique automobiles."

WILLIAM T. SILFVAST ("Metal-Vapor Lasers") is a member of the technical staff of the Bell Telephone Laboratories. His higher education was at the University of Utah, where he obtained a bachelor's degree in physics and in mathematics in 1961 and a Ph.D. in physics in 1965. After a year as a research associate at the university and a year at the Clarendon Laboratory of the University of Oxford he joined Bell Laboratories. "My outside interests," he writes, "include painting, music and woodworking. I am also in the process of restoring an old imported sports car. I like most sports and in particular I enjoy teaching them to my two young boys. Our family loves to travel and we find that we are often attracted to many of the relatively unpopulated areas of New England."

FRANCO PACINI and MARTIN J. REES ("Rotation in High-Energy Astrophysics") are respectively with the Space Astrophysics Laboratory of the National Research Council of Italy and the Astronomy Centre of the University of Sussex. Pacini received his doctorate in physics from the University of Rome, was a postdoctoral fellow at the Institute of Astrophysics in Paris and spent three and a half years at Cornell University. In addition to his work with the National Research Council he is adjunct professor of space physics at the University of Rome. He says he is "probably one of the few Italians who share their lives with an Alaskan Malamute dog." Rees, who is professor of astronomy at Sussex, obtained his master's degree in mathematics and his Ph.D. in astrophysics from the University of Cambridge, where he later worked at the Institute of Theoretical Astronomy before going to Sussex.

I. BERNARD COHEN, who in this issue reviews *Science, Technology and Society in Seventeenth-Century England,* by Robert K. Merton, is professor of the history of science at Harvard University.

Established 1845 SCIENTIFIC AMERICAN February 1973 Volume 228 Number 2

Reconnaissance and Arms Control

Reconnaissance satellites are the chief means relied on by the U.S. and the U.S.S.R. to verify each other's compliance with the SALT I accords. What bearing will they and related systems have on SALT II?

by Ted Greenwood

The two major arms-control agreements signed last May in Moscow as a result of the first round of the strategic-arms-limitation talks (SALT I) between the U.S. and the U.S.S.R. both incorporate sections designed to deal with the problem of verification. For example, Article XII of the treaty on the limitation of anti-ballistic-missile (ABM) systems states: "1. For the purpose of providing assurance of compliance with the provisions of this Treaty, each Party shall use national technical means of verification at its disposal.... 2. Each Party undertakes not to interfere with the national technical means of verification of the other Party.... 3. Each Party undertakes not to use deliberate concealment measures which impede verification by national technical means of compliance with the provisions of this Treaty...." Article V of the SALT I interim agreement "on certain measures with respect to the limitation of strategic offensive arms" uses virtually identical language to make the same three points.

What are the "national technical means of verification" mentioned in these documents? How are these systems now being used by each side to unilaterally monitor the other side's compliance or noncompliance with the SALT I accords? What bearing does this mutual capability have on further arms-control agreements, in particular on those that might emerge from the current SALT II negotiations?

In attempting to answer such questions this article will of necessity focus primarily on the subject of observation satellites, since the main restrictions imposed by both of the SALT I agreements can be, and undoubtedly are being, monitored largely by means of sensors carried on board such orbiting photoreconnaissance systems. The primary SALT I restrictions limit each side's ABM missile launchers to 200 at two widely separated sites, impose numerical ceilings on both land-based intercontinental ballistic missiles (ICBM's) and submarine-launched ballistic missiles (SLBM's), regulate the replacement of old ICBM's and SLBM's with new SLBM's and prohibit the exchange of old ICBM's or light ICBM's for heavy ICBM's.

In addition the ABM treaty bans the testing of new types of ABM systems or of other systems or components "in an ABM mode." The monitoring of activities of this type involves not only observation satellites but also other methods specifically suited for the surveillance of missile launchings. Hence systems such as advanced land-based radars, early-warning satellites and shipboard tracking sensors are the second major concern of this article. Such systems are of particular importance to any discussion of the prospect that SALT II will result in meaningful qualitative restrictions of strategic offensive missiles.

In spite of the official secrecy that almost completely hides from public view the observation-satellite programs of both the U.S. and the U.S.S.R., it is possible to make some general observations about the theoretical capabilities of such systems and about their actual performance. For example, it is known that because of the fundamental electromagnetic properties of the earth's atmosphere, satellite-borne sensors are restricted to three spectral "windows": the visible-light wavelengths, a broad infrared band centered on a wavelength of eight microns and certain radar wavelengths. Over the years, as larger booster rockets have made it possible to put heavier instrument packages into orbit, there has been an increase in the number of different sensors and in their resolution and film capacity. At the same time the associated communications systems have also become more sophisticated and their data-transmission rate has increased.

The most useful measure of the quality of an airplane or satellite photoreconnaissance system is ground resolution, a value that is equivalent to the smallest object that can be distinguished on the ground with good contrast. For a camera pointing vertically downward the expression for ground resolution, G, is given by the equation $G = A/300\ FR$, where A is the altitude of the airplane or the satellite in feet, F is the focal length of the camera in feet, R is the joint resolution of the film and optics in lines per millimeter and 300 is a numerical constant characteristic of the units used for the other quantities.

To cite a recent example of space photography, the panoramic camera built by the Itek Corporation for the last three Apollo missions to the moon had a focal length of two feet and photographed the lunar surface from an altitude of approximately 60 miles. The resulting photographs had a ground resolution of about three feet [*see illustration on pages 24 and 25*]. These numbers, inserted in the equation given above, suggest an optical resolution for the system of better than 180 lines per millimeter.

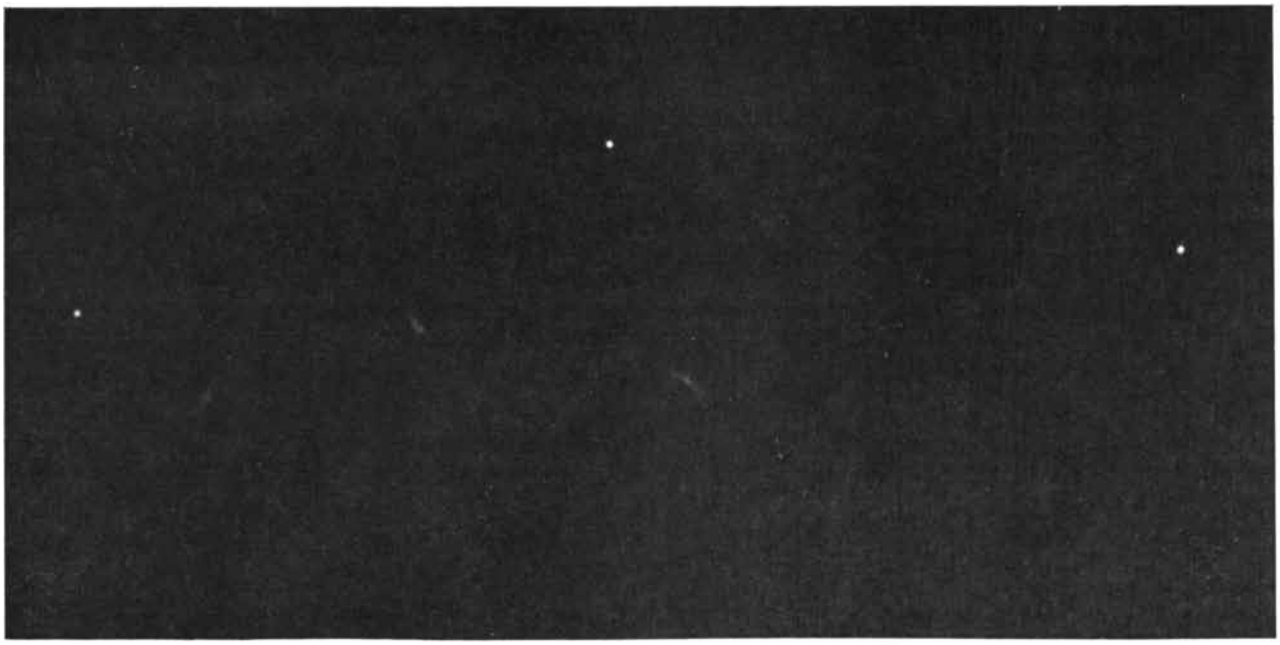
By way of comparison, it has been reported unofficially that the high-resolution cameras used in the newest generation of observation satellites have a focal length of more than eight feet. Taking eight feet as the focal length and 180 lines per millimeter as the optical resolution, one can calculate that the ground resolution of such a system from a nominal altitude of, say, 100 miles would be 1.2 feet. Actually there is reason to believe that the resolution of the most advanced systems now in service may be even better; as early as 1960 experts in the field were talking about prospective U.S. Air Force cameras with focal lengths on the order of 20 feet. Furthermore, films with a resolution considerably better than 180 lines per millimeter are currently available. A compromise must be made, however, between film resolution and film speed, and rather fast film is needed for satellite photography.

Another factor that must be taken into consideration is the aperture of the camera's optics. Diffraction effects place an upper limit on the optical resolution of any lens system (a limit never attainable in practice). The limit is given by the ratio $.82d/\lambda F$, where d is the aperture in feet and λ is the wavelength in millimeters of the radiation to which the system is sensitive. For a camera with a focal length of eight feet an aperture of one foot would result in a diffraction-limited resolution of 180 lines per millimeter, assuming that one is working with visible light at a wavelength of .55 micrometer.

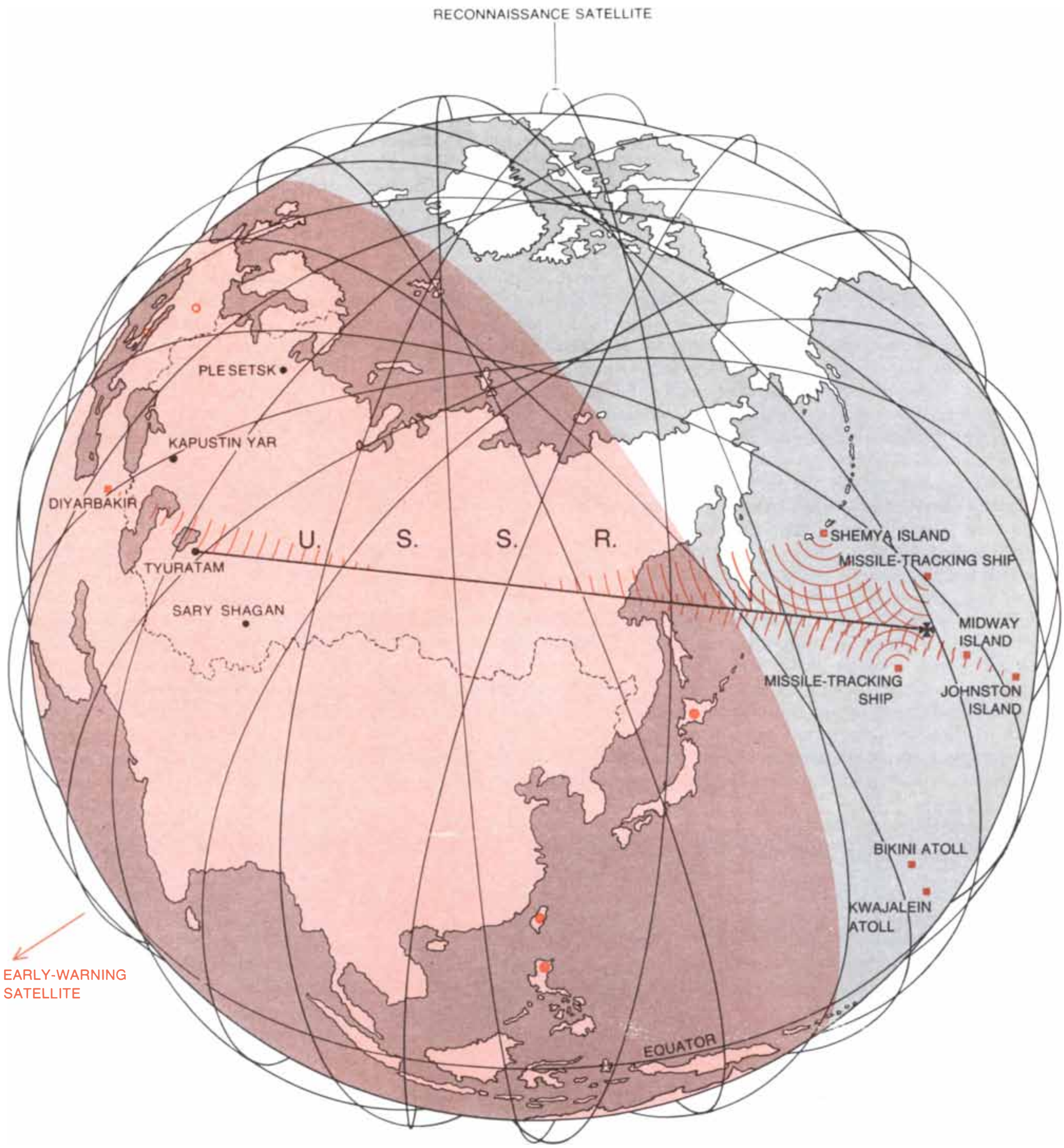
In practice a much larger aperture would be employed in order to ensure that the optical system is not limited by diffraction. Therefore a camera with a focal length of eight feet and an aperture of three feet would (assuming that it uses film with a resolution of 180 lines per millimeter) provide a ground resolution of 1.6 feet from an altitude of 100 miles. With the same film a camera having a focal length of 20 feet and an aperture of five feet would yield a ground resolution of .7 foot. To achieve a ground resolution of .5 foot the same camera would require film with a resolution of about 330 lines per millimeter.

These simple calculations are based on the assumption that the overall resolution of the system is poorer than the poorest contributing factor, which may be either the diffraction-limited resolution of the optics or the line resolution of the film. Moreover, the results do not take into account various operational degradations, nor do they involve the intrinsic limitation on ground resolution presented by the scattering of light in the atmosphere. The latter limitation amounts to a few inches for an altitude of 100 miles. Nonetheless, it seems clear that the latest U.S. photoreconnaissance satellites are close to achieving resolutions that are limited only by atmospheric effects.

Ground resolution is not the only relevant measure of picture quality. It characterizes an image only at the threshold of what can be barely distinguished. The color and light contrast of the objects being photographed strongly affect their distinguishability. The angle of the sun and the intrinsic reflectivity of an object are also important, since these factors influence the object's apparent brightness. In general the task of identifying an object in a photograph is always more difficult than simply locating it. On the other hand, extended objects such as roads or rail lines can often be distinguished even though their dimensions



RUSSIAN MULTIPLE-WARHEAD TEST was photographed from a U.S. Air Force plane in the vicinity of the projected impact point somewhere in the Pacific Ocean. The three separate reentry vehicles, which were launched by a single SS-9 missile from the Russian missile-testing center at Tyuratam, east of the Aral Sea, are equipped with protective head shields that glow as the vehicles reenter the atmosphere. As a result the vehicles (which for tests of this type bear dummy nuclear warheads) are visible as bright dots in the photograph. The three fainter streaks at lower left were made by fragments of the booster rocket burning up on reentry into the atmosphere. U.S. radars routinely monitor Russian missile tests near both ends of their trajectory, after which specially equipped aircraft and vessels are dispatched to make such photographs. This picture was released by the Department of Defense in 1971.



RECONNAISSANCE AND SURVEILLANCE SYSTEMS currently used by the U.S. to verify Russian compliance with the terms of the SALT I agreements are represented schematically in this illustration. U.S. photoreconnaissance satellites are typically launched (from Vandenberg Air Force Base in California) into a near-polar elliptical orbit with an orbital period of approximately 90 minutes and a perigee (lowest point) on the order of 100 miles. The latest, fourth-generation U.S. observation satellite, unofficially called Big Bird, combines the separate functions of area-surveillance photography and close-look photography and hence is required to stay aloft for a much longer period than earlier close-look satellites; orbital times to date have averaged about seven weeks. The orbit of such a satellite remains essentially fixed in space while the earth rotates, with the result that to an earth-based observer the satellite appears to move westward on each successive orbit. Hence most of the earth's surface passes under the orbital path of the satellite. U.S. early-warning satellites, in contrast, are typically launched into near-equatorial, near-synchronous "parking" orbits at altitudes of about 22,300 miles. Two such satellites, launched into identical "figure eight" orbits at the same fixed longitude over the Indian Ocean but lagging each other by 12 hours, can provide continuous infrared coverage of most of the U.S.S.R. and all of China (*light colored area*). The value of early-warning satellites from the point of view of arms control is that they are also capable of monitoring missile tests. The black dots indicate the locations of the major Russian missile-testing launch centers; the black line shows a typical trajectory for a Russian long-range missile test. Also shown are two types of U.S. radar used to monitor Russian missile tests: over-the-horizon transmitters (*colored dots*) and receivers (*open colored circles*), and conventional, or line-of-sight, radars (*colored squares*).

may be far below the calculated ground resolution.

Two different techniques are currently employed by the U.S. to retrieve photoreconnaissance information from space. The first is to return the film itself for processing on the ground. This method preserves all the information that is recorded on the film and is therefore employed for very-high-resolution photographs. The exposed film is ejected from the satellite and is returned to the earth in a special reentry package, which is caught in midair by specially equipped aircraft as it floats by parachute through the latter part of its descent [*see illustration on page 20*].

For lower-resolution pictures the film can be developed on the satellite and scanned by a television camera or a laser system. This information is temporarily stored and then transmitted to the earth when the satellite passes over one of several ground and shipboard stations around the world. Clearly in order to transmit all the necessary information during the short time in which the satellite is in view of these stations a communications system with a high data-transmission rate is needed.

Once a film pack is returned to the earth it must be developed to maximum advantage and then interpreted. The transmitted pictures must also be recreated and then interpreted. With a ground resolution measured in inches or feet and with a film resolution measured in the hundreds of lines per millimeter, it is clear that by the time enlargements of the photographs are made many man-hours must be expended to interpret them. Indeed, photographic interpretation must be a very large enterprise today both in the U.S. and in the U.S.S.R. It is also no wonder that systems are used to make photographs at less than maximum resolution in order to identify interesting targets for further photography at higher resolution.

In addition to visible-light cameras infrared sensors are used in reconnaissance satellites. Since all bodies at terrestrial temperatures radiate energy that is predominantly in the infrared part of the spectrum, infrared photography is not dependent on the sun for illumination. It is therefore useful for night applications or in polar regions. (Low-light systems have been developed and may also be in service for these purposes.) Perhaps the greatest significance of infrared sensors is their ability to detect things (such as missile silos) that are underground or may be camouflaged. As long as the ground immediately around the object is at a different temperature or has emission characteristics different from those of the surrounding terrain, an underground silo will stand out in an infrared picture. For this reason multispectral photography, a technique for making pictures simultaneously at several wavelengths, is increasingly employed in observation satellites. Although longer-wavelength infrared photography would require much larger optics to achieve resolutions comparable to those attainable with visible-light systems, such high resolution is probably not needed to accomplish the tasks assigned to the infrared sensors.

One of the major obstacles for satellite reconnaissance is, of course, cloud cover. Certain locations, for example the location of Moscow, are rarely free of clouds. The coordination of observation-satellite launchings with information received from weather satellites can overcome this problem to some extent. Moreover, observation satellites currently in service probably have the capability of changing their orbit to take advantage of breaks in the cloud cover.

Observation satellites contribute to the verification of the SALT agreements in a variety of ways. By photographing missile test sites they help to identify new missile systems, to detect changes in operational procedure that may suggest a change in hardware and to monitor testing programs. They can watch industrial facilities, including shipyards for the construction of submarines and plants for the assembly of missiles. Intermittent information on critical aspects of transportation networks can be obtained. The progress of construction of missile silos, ABM radars or ABM launcher sites can be monitored. Wide-area surveys can be made to determine if any activity is underway that violates the agreements.

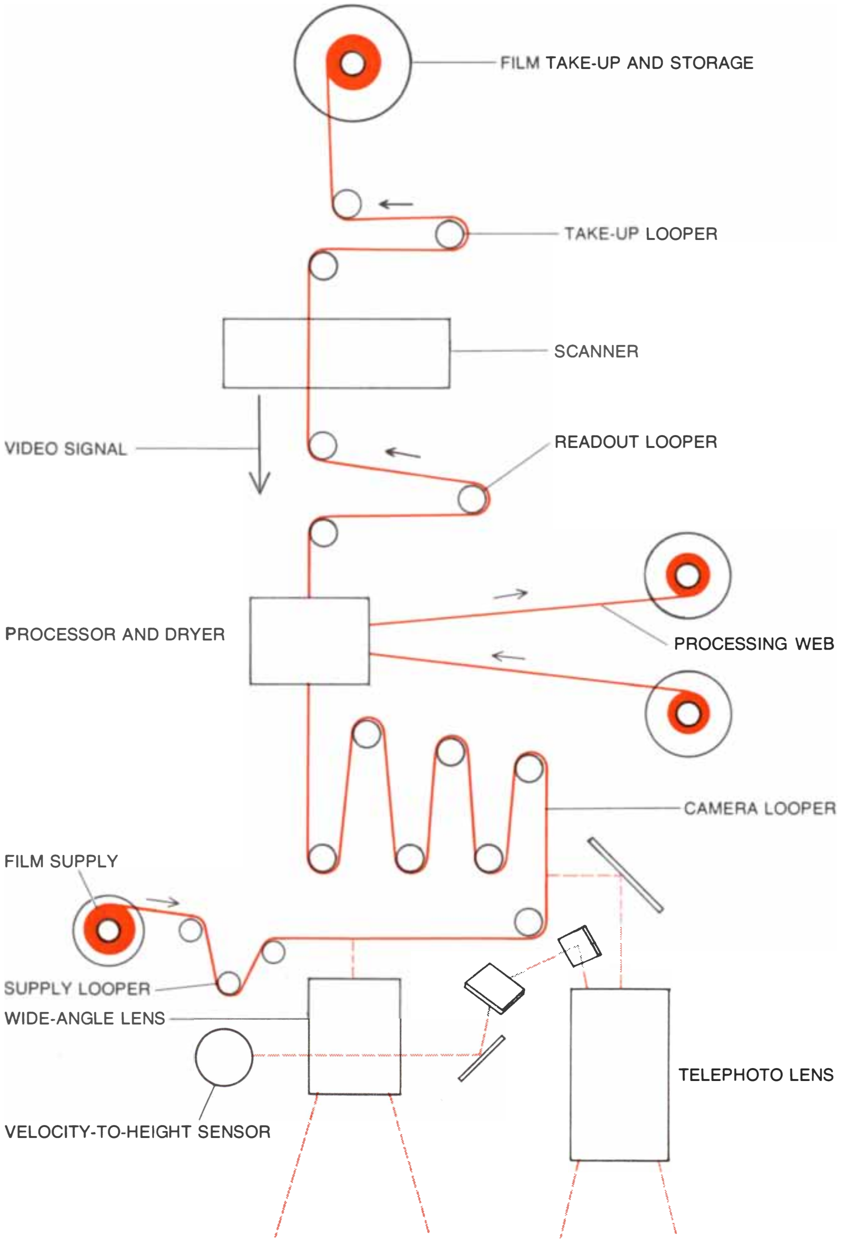
Although the possibility of artificial earth satellites and their potential for military reconnaissance were recognized immediately after World War II, not until the mid-1950's was the future availability of rocket boosters assured and a satellite program actually initiated. On March 16, 1955, under the sponsorship of the Central Intelligence Agency, the U.S. Air Force issued a formal operational requirement for a strategic satellite system. After a year-long competition the Air Force selected the Lockheed Aircraft Corporation to develop a self-powered satellite vehicle, later named the Agena. This vehicle (along with several later models) was for many years the workhorse of the U.S. observation-satellite program.

The U.S. did not wait for the availability of satellites, however, before beginning a strategic-reconnaissance program. By 1956 U-2 reconnaissance aircraft were making their first flights over the U.S.S.R. Their photographic information supplemented the radar data on Russian missile tests, provided the basis for the gradual downgrading of estimates of the size of the Russian bomber force and eventually convinced most informed observers that the Russian ICBM buildup was proceeding at a much slower pace than had been anticipated. In spite of the considerable usefulness of the U-2 flights, the aircraft had two weaknesses as a sensor platform. First, the limited range of the plane, the endurance of the pilot and the provocative nature of the mission imposed a severe limitation on the area of the U.S.S.R. that could be photographed. As a result of this limitation the lack of U-2 evidence for large numbers of deployed Russian ICBM's was not conclusive proof that only a few existed. Not until reconnaissance satellites had provided much greater coverage of the landmass of the U.S.S.R. could the U.S. intelligence community be certain that missiles had not been deployed and remained undetected. The second weakness of the airplane was its vulnerability. Eventually the Russians developed an antiaircraft missile that could hit the U-2 at 70,000 feet. When Francis Gary Powers was shot down in May, 1960, the U-2 overflights ended, except for certain minor incursions. Satellites, on the other hand, were known to be invulnerable, and would remain so for the then foreseeable future.

Among the many technical problems that had to be solved in order to develop an operational reconnaissance satellite were the stabilization and orientation of the spacecraft, the design and production of light cameras with a long focal length and a large aperture, and the recovery of the data. The possibility of using a television camera was considered very early but was rejected because the desired resolution was not attainable with available technology. In 1957 the decision was made to pursue two parallel approaches: first, direct recovery of a film package, and second, on-board developing and scanning of film followed by radio transmission to ground stations. Techniques for direct recovery were developed using Discoverer satellites as test vehicles. The first successful reentry package was recovered from *Discoverer*

13 on August 11, 1960. Either this package or the one recovered from *Discoverer 14* the following week probably yielded the first satellite photographs of the U.S.S.R. That was just three months after the last U-2 overflight of the U.S.S.R. The first successful radio-transmission observation satellite was *Samos II,* which was put in a polar orbit varying in altitude between 300 and 350 miles on January 31, 1961; this spacecraft carried between 300 and 400 pounds of instruments.

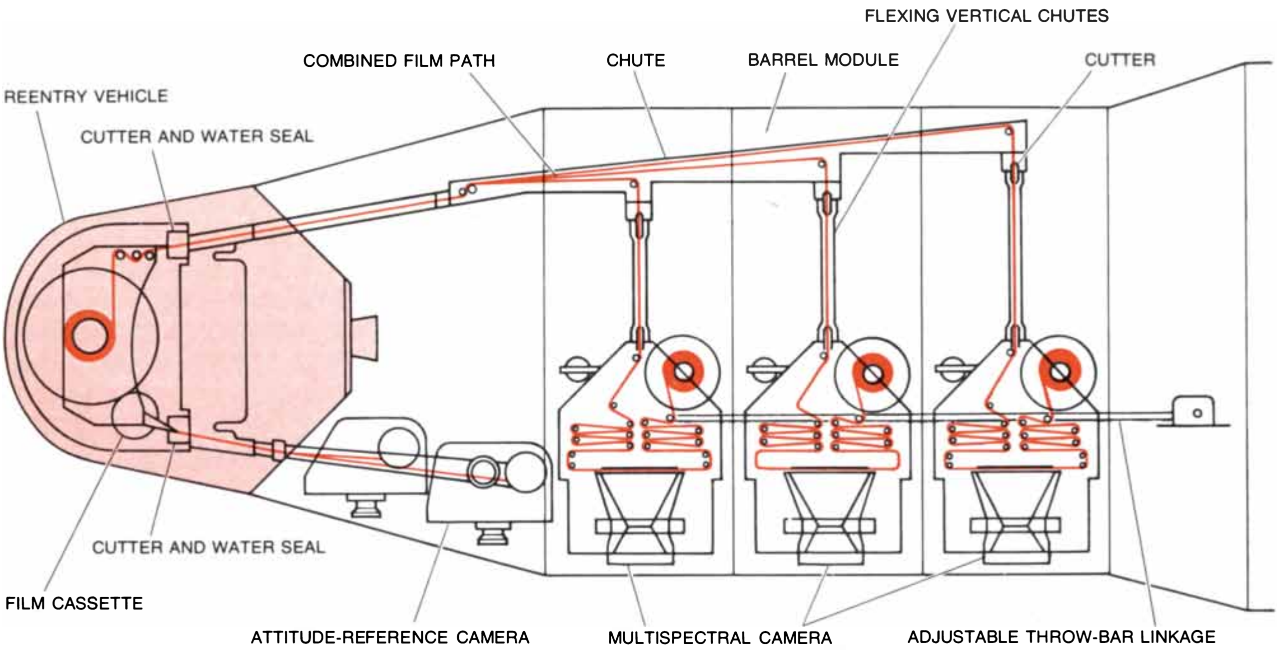
These early successes almost certainly were responsible for the final laying to rest of the myth of the "missile gap." On February 6, 1961, Secretary of Defense Robert S. McNamara told reporters at an informal, off-the-record briefing that a study had yielded no evidence of such a gap. A White House spokesman denied the report, saying that the studies were not yet complete. Intelligence sources, however, soon began to reduce their estimates of the number of Russian missiles. By September, 1961, after the successful recovery of additional Discoverer capsules and with time to analyze the *Samos II* data, the number of deployed Russian ICBM's was reportedly put at 14.



AREA-SURVEILLANCE SYSTEM similar to the system used in the early U.S. Samos reconnaissance satellites is represented here by a schematic diagram of the interior mechanism of a Lunar Orbiter, the spacecraft used to make a photographic survey of the surface of the moon preparatory to selecting the Apollo landing sites. In both systems the film is exposed and developed by the optical and film-processing elements of the camera subsystem; the developed photographs are then scanned by a television camera, converting the picture into electrical signals for radio transmission back to earth. Cameras for both Lunar Orbiter and Samos were built by the Eastman Kodak Company; film scanners for both spacecraft were built by CBS Laboratories, a division of the Columbia Broadcasting System.

The descendants of the early Samos satellites have been comparatively small radio-transmission observation satellites whose sensors and orbital characteristics are chosen to maximize their degree of coverage. They stay in orbit for three to four weeks. Although their perigee (lowest orbital point) is about 100 miles above the earth's surface, the ellipticity of their orbit greatly reduces their atmospheric drag, thereby increasing their lifetime. With an inclination of 80 to 92 degrees with respect to the Equator, they provide virtually full coverage of the U.S.S.R. and complete coverage of China. These "area surveillance" satellites provide low-resolution coverage of wide regions, and the radio transmission of data makes possible the rapid recovery of intelligence information. Particular areas of interest are identified on the basis of this information and higher-resolution pictures are later made by another satellite with a recoverable film capsule. By early 1962 the size and weight of the camera system had been reduced sufficiently to allow the area-surveillance satellites to be launched by the Thor/Agena booster rather than by the more powerful Atlas/Agena booster used earlier.

The use of the Thrust-Augmented-Thor/Agena D, beginning in May, 1963, indicates that a new generation of satellites was then introduced. The greater throw weight of this booster rocket enabled the satellite to carry a larger camera and more of a consumable payload (including film), resulting in an increase in useful lifetime. These satellites were launched at roughly one-month intervals beginning in the middle of 1963. Until the end of 1965, however, two or more area-surveillance satellites were frequently in orbit at the same time, and occasional delays in the regular launch schedule can be identified. This suggests that problems of reliability necessitated the replacement of satellites that had failed before the completion of their mission and also that problems had sometimes occurred at ground level. Since 1966 one of these satellites has been in orbit for almost half the days of



CLOSE-LOOK SYSTEM similar to the system used in the U.S. second-generation Discoverer reconnaissance satellites is represented here by a schematic diagram of the interior mechanism of a proposed nonmilitary satellite. The scheme shown was submitted by the General Electric Company to the National Aeronautics and Space Administration in 1969 for possible use in NASA's earth-resources survey program. In both systems exposed film is fed from several cameras into a recoverable capsule, which is ejected from the satellite and returned to earth for processing. Since this method preserves all the information on the film, it is preferable for high-resolution photography. The recoverable-film system for the Discoverer reconnaissance satellites was produced by General Electric.

the year, and there has been hardly any overlap. One can therefore probably fix the beginning of 1966 as the date of a fully operational status for the area-surveillance satellites.

The first test launch employing the even more powerful Long-Tank-Thrust-Augmented-Thor/Agena was in August, 1966, and that booster was introduced into regular service in May, 1967. The use of this booster also marks the arrival of a heavier, third-generation satellite. It was probably equipped with a camera of longer focal length, a larger film supply, an infrared optical system and a new transmission system with an increased data rate.

The task of these satellites has been to survey wide areas with sensors of moderate resolution and to reveal targets that merit a closer look at higher resolution. In order to provide that closer look, a different type of satellite is used.

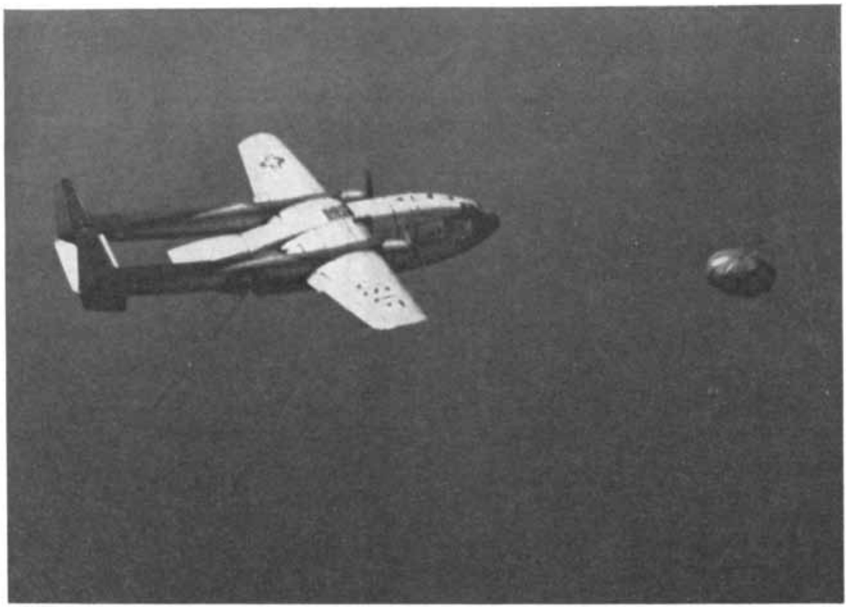
The descendants of the Discoverer recoverable-capsule satellites are the close-look satellites. They are heavier than the area-surveillance satellites, reflecting the fact that they carry a camera with a longer focal length and a wider aperture. They are also in a lower orbit, with a perigee of typically about 80 miles, in order to maximize resolution. With this more powerful telescopic system interesting targets identified by an earlier area-surveillance satellite can be rephotographed and examined more closely. To minimize information losses introduced by electronic data storage and transmission, the close-look satellites send their film packs back to the earth in a reentry capsule for midair recovery by aircraft.

The first launching of a close-look satellite appears to have been on April 26, 1962, when a Thor/Agena booster put into orbit a satellite carrying a recovery capsule designated *E*-6. Three days later the capsule and its film were recovered. Judging from the frequency of launchings, this program seems to have achieved operational status by the middle of 1963. With their shorter lifetime in orbit, the close-look satellites had less stringent reliability requirements than the area-surveillance models and were therefore able to reach full operational status more quickly. By 1964 the Atlas/Agena had been introduced as the program's booster, ensuring that heavier satellites with improved capabilities could be placed in orbit. These second-generation high-resolution satellites were launched about once a month and remained in orbit for three to five days before sending their film package back to the earth.

The Titan 3B began to be used for test launches in July, 1966, and came into regular service in August, 1967. With the introduction of this still larger booster the lifetime of the close-look satellites began to increase until by 1968 they remained in orbit for a period averaging some two weeks. This clearly shows that a third-generation satellite had been introduced with a much greater film capacity and the ability to raise its orbit in order to avoid early burnup. It has been suggested that the new satellite could alter its orbit to take advantage of breaks in cloud cover; moreover, it seems likely that several new types of sensor were included in the satellites. Infrared sensors and multispectral photography appear to be the most likely candidates because of their ability to discover and penetrate camouflage and, in the case of the infrared sensors, to operate in the dark. An accurate mapping camera for the purpose of pinpointing the location of strategic targets in the U.S.S.R. may also have been included.

In the past 18 months or so an entirely new fourth generation of observation satellites has been introduced and is now reaching operational status. This satellite, unofficially called Big Bird, weighs more than 20,000 pounds, which makes it much heavier than any previous observation satellite. The spacecraft itself is a modified Agena rocket 10 feet in diameter and 50 feet long. It is launched by the powerful Titan 3D booster.

The extra weight and size of this sys-



MIDAIR RECOVERY of a film capsule from a Discoverer-type reconnaissance satellite is accomplished by means of specially equipped aircraft as the capsule floats by parachute through the latter part of its descent. This photograph, released by the Air Force in 1961, shows a C-119 transport aircraft "as it approaches a Discoverer capsule... somewhere over the Pacific Ocean." At present larger C-130 aircraft are used routinely to recover the much heavier film capsules dropped by the latest-model close-look reconnaissance satellites.

tem result from joining the separate functions of area-surveillance and close-look photography into the one satellite. Big Bird is reported to carry an area-surveillance camera made by the Eastman Kodak Company and an on-board film processor and scanner. The resulting data are reportedly transmitted by means of a new 20-foot unfurlable antenna, which would represent an increase in capacity by a factor of 16 over the older five-foot antennas. In the past, several months would go by before a close-look satellite could be launched to rephotograph an area of interest identified by a low-resolution photograph and its film pack could be recovered. Now, however, Big Bird can be directed to turn on its high-resolution camera (made by the Perkin-Elmer Corporation) during a subsequent pass. Film from this camera, said to have a resolution of less than one foot from an altitude of 100 miles, is returned in one of several recovery capsules. The delay time should now be cut to several weeks.

This dual capability requires that Big Bird remain aloft for a much longer period than earlier close-look satellites. In order to accomplish this result the satellite is placed in a higher and more elliptical orbit. The orbital characteristics of the first Big Bird, launched on June 15, 1971, were a perigee of 111 miles and an apogee of 180 miles. To compensate for this higher altitude and to improve resolution both the focal length and the aperture of the high-power camera had to be increased over those of earlier models. To further increase the satellite's lifetime it has been equipped with an on-board rocket to raise its orbit and prevent early burnup. The times in orbit for the first three satellites were respectively 52 days, 40 days and 68 days. The fourth launching was on October 10, 1972.

Certainly since 1962, and probably earlier, the U.S. has had detailed information on the number and location of Russian strategic missiles. In 1967 President Johnson told a meeting of educators in Tennessee that satellite reconnaissance was worth 10 times the money the U.S. had spent in space. "I know how many missiles the enemy has," Johnson said. He suggested that this knowledge had prevented the country from harboring fears that otherwise might have arisen. From the late 1960's to the present the Department of Defense has regularly published information on the level of Russian ICBM, SLBM and ABM deployment. Both Congress and the public have come to expect such information as part of the Department's budget justification, and in recent years both the Administration and its critics have used the published figures to support their own arguments.

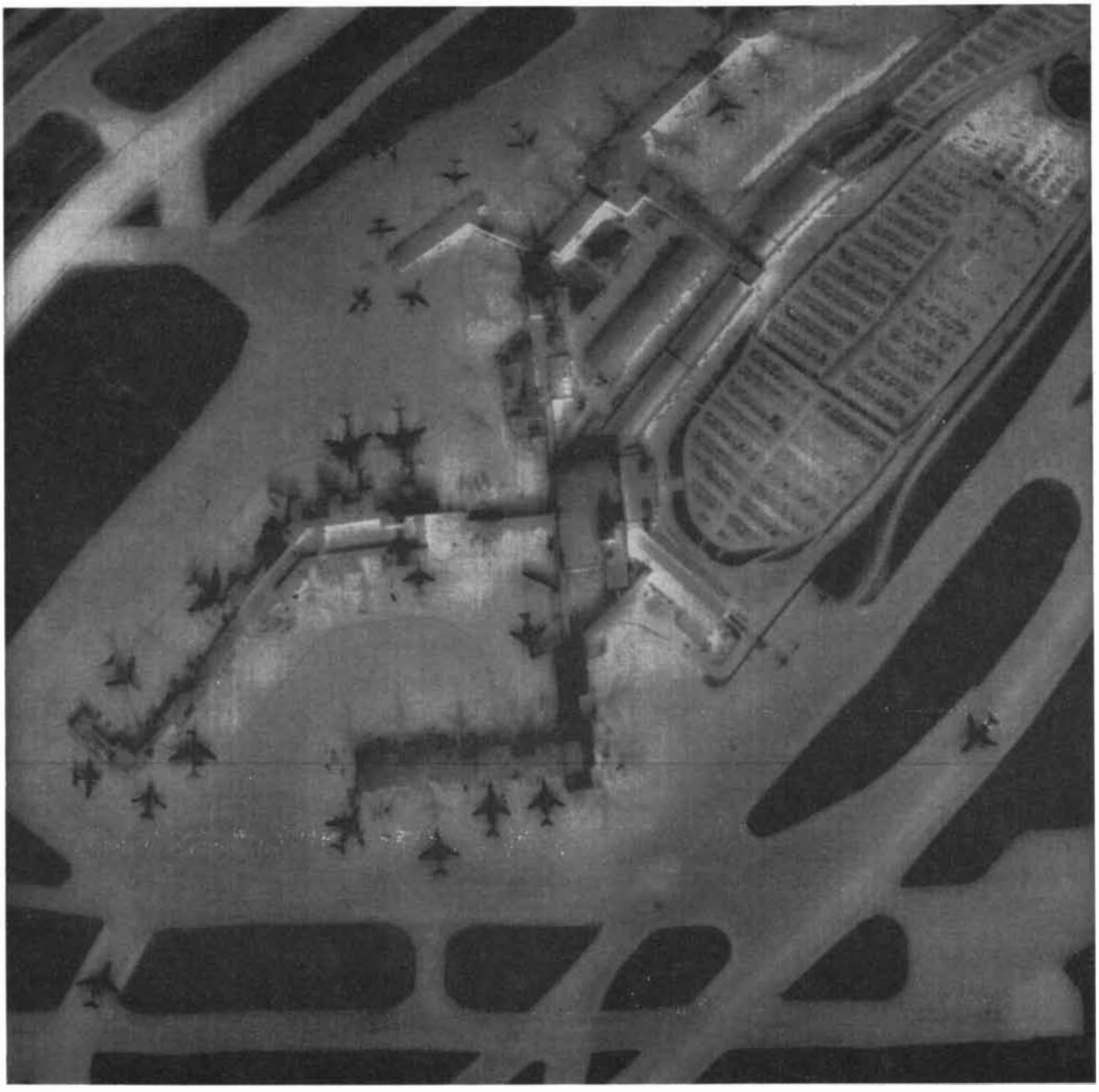
The major task assigned by SALT I to these now protected "national technical means of verification" is to monitor the quantitative limits imposed on the number of ABM launchers, large radars, ICBM's and SLBM's. This is the arms-control task for which observation satellites are best suited. From satellite photographs intelligence analysts are able to monitor silo construction and the transport of missiles to their deployment sites. With multispectral and infrared photography they can detect or penetrate camouflage and monitor nighttime activity. Submarine shipyards are observed on a routine basis to monitor new construction. Similarly, ABM launchers and radar deployments can be observed. There seems to be no doubt that these aspects of the SALT I agreements can be verified by satellite reconnaissance with a high degree of confidence, as long as the Russians live up to their pledge "not to use deliberate concealment measures which impede verification by national technical means."

The more important question, however, is whether or not the Russians could find ways clandestinely to circumvent the restrictions and thereby achieve a military or political advantage over the U.S. Although the actual deployment of prohibited weaponry secretly seems quite impossible, could the Russians simply manufacture additional missiles or radars and then abrogate the treaty at a time of their own choosing and deploy them on a time scale too short for a U.S. response? A full discussion of this problem would have to include an assessment of the internal pressures that might lead the U.S.S.R. to such actions, the possible international political repercussions and an analysis of what the Russians would have to do that would make any difference either militarily or politically.

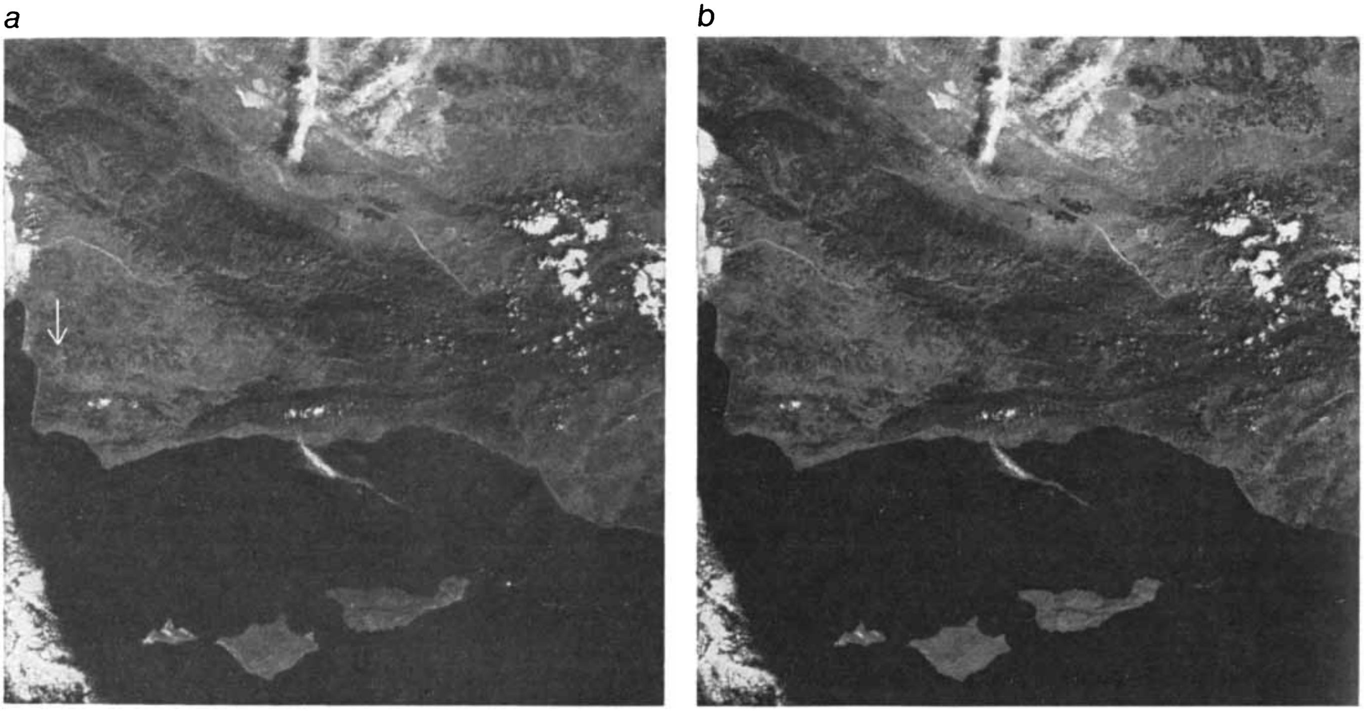
Although such considerations lie outside this discussion, one can still go part of the way toward dealing with the problem. The continuous monitoring of Russian transportation networks, power-generation plants and manufacturing facilities by observation satellites would make it unlikely that the U.S. could not detect such clandestine activity in time to react in some way. The decrease in delay time provided by Big Bird between the first indication of suspicious activity in an area-surveillance photograph and further photography with a high-resolution camera will certainly reduce the uncertainties involved. Although photography cannot penetrate buildings, infrared and multispectral techniques can often reveal a great deal about activities inside, particularly since activities of special interest may be indicated by changes in standard operating procedures.

Cloud cover remains an impediment to observational activities, but with longer orbital lifetimes and the capability of changing orbit the newer satellites are less constrained than their predecessors. For the clarification of ambiguities both the unilateral techniques of electronic and communications intelligence and the bilateral apparatus of the Standing Consultative Commission (also set up by the SALT I agreements) may be useful. To summarize, the very size and complexity of construction and industrial activity required to build and deploy modern strategic weapons, combined with the breadth of coverage, resolution and multispectral aspects of observation satellites, guarantee that the agreements of SALT I can be verified unilaterally with high confidence. The cost of hardware and manpower for photointerpretation, however, will remain high.

If a future agreement were to impose qualitative restrictions on strategic systems, the U.S. would have to rely on other verification techniques to augment the observation satellites. The usefulness in this regard of systems that monitor missile tests will be considered below. Observation satellites would also, however, have a role to play in monitoring such agreements. Any qualitative improvement in missile weaponry, whether it is new guidance systems, multiple warheads, improved ABM radar, new SLBM's or advanced ICBM's, must be reflected in changes from former manufacturing and testing procedures and equipment. Any new hardware must be delivered to an operational site and



INFRARED PHOTOGRAPH of an airport in Texas was made from an aircraft flying at an altitude of 2,000 feet. Since infrared photography is not dependent on the sun for illumination, it is used in reconnaissance satellites for night applications or in polar regions. The ability of infrared sensors to detect the heat from missile exhausts is suggested by the bright images of the jet exhausts from the aircraft on the runway at lower right. This unusual unclassified photograph was supplied by Texas Instruments Incorporated.



MULTISPECTRAL PHOTOGRAPHY, a technique for making pictures simultaneously at several wavelengths, is increasingly employed in reconnaissance satellites. The four sample photographs shown on these two pages were made by a multispectral scanner on board NASA's Earth Resources Technology Satellite; the line-scanning device employed in this case uses an oscillating mirror to simultaneously record the terrain passing under the spacecraft in four spectral bands: .5 to .6 micrometer (*a*), .6 to .7 micrometer (*b*), .7 to .8 micrometer (*c*) and .8 micrometer to 1.1 micrometers (*d*). The electronic signals produced by this television system are

installed. These activities can be monitored by satellites.

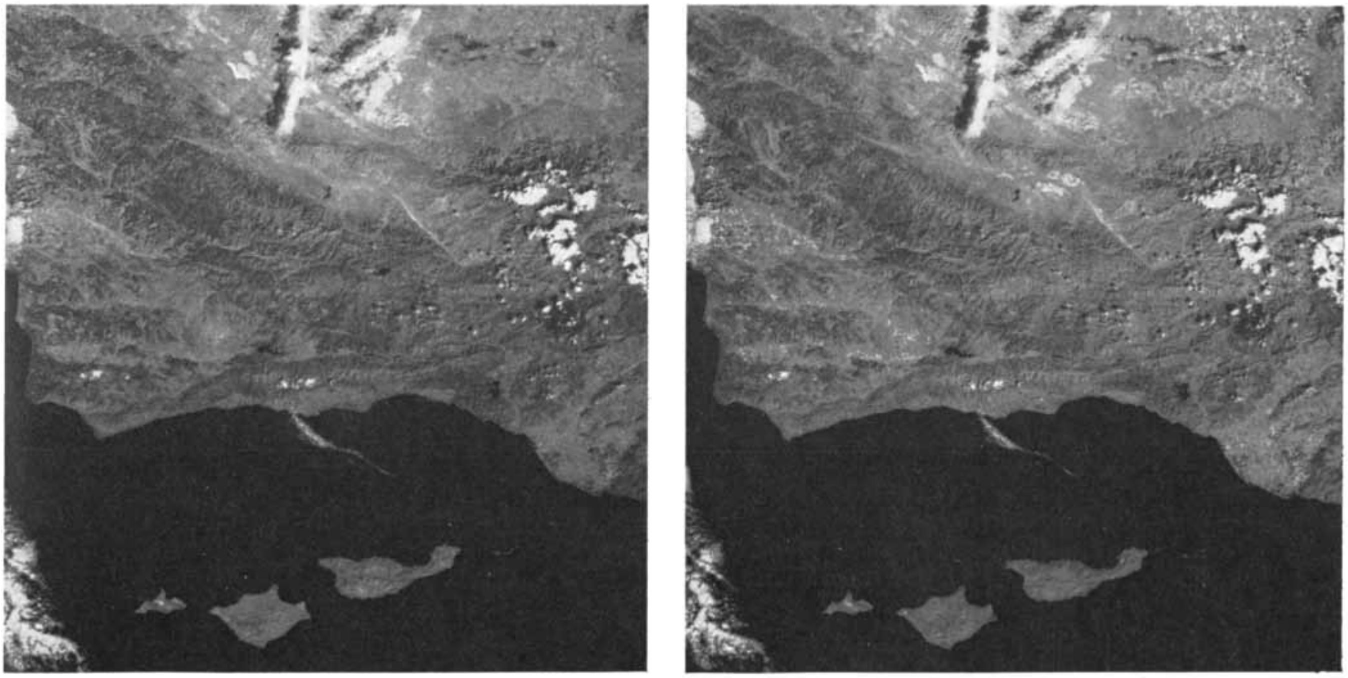
The surveillance systems that would be useful in the verification of an agreement imposing qualitative restrictions on ballistic missiles include certain land-based line-of-sight radars, over-the-horizon radars, satellite systems and particularly shipboard sensors for terminal observations. As early as the summer of 1955 a U.S. radar at Samsun in Turkey was tracking missile tests from the Russian launch site at Kapustin Yar, northwest of the Caspian Sea. As a result of this monitoring the launching of *Sputnik I* in October, 1957, and the Russian ICBM tests of the same year came as no surprise to the U.S. intelligence community and Government officials with access to such data. They were well aware of the Russian capabilities in this area. By late 1963 or early 1964 a longer-range radar had been installed in Diyarbakir in Turkey, bringing into view missiles launched from the newer Russian test center at Tyuratam, east of the Aral Sea.

Several fixed land-based radars are also available to observe Russian tests near the end of their flights. One of these has been operational on Shemya Island, far out in the Aleutian chain, since at least 1959. This radar can track and provide data on the Russian reentry vehicles that impact either in the test area on Kamchatka Peninsula or in the North Pacific area northwest of the Midway Islands. For the longer-range tests that end in the Pacific southwest of Johnston Island several radars that have been installed for other purposes at the Midway Islands, Bikini Atoll, Kwajalein Atoll and Johnston Island can be employed.

Unlike conventional radar, over-the-horizon, or OTH, radar is not restricted in its range by the curvature of the earth. By reflection from the ionosphere OTH radar can penetrate to great distances, making possible the detection of missiles soon after they are launched. The currently deployed "forward scatter" OTH radar detects the disturbances in the ionosphere caused by the ionized jet of gas emanating from a rocket's motor. Since each type of missile disturbs the ionosphere somewhat differently, a detected missile can be identified by its characteristic OTH signature. In the currently operational system three transmitters are deployed in Taiwan, Japan and the Philippines. These transmitters are matched with corresponding receivers in Italy, Germany and another European country. Although the system was originally intended as an early-warning system for a massive missile attack, it has detected a high percentage of the known single events since 1968. All long-range missiles fired from test sites in the U.S.S.R. are detectable.

Parallel to the observation-satellite developments of the late 1950's and early 1960's there was an expensive and unsuccessful program to develop a satellite early-warning system. For years this program was plagued by unreliabilities in its hardware and by the inability of its infrared sensors to distinguish between rocket-exhaust plumes and sunlight reflected from high clouds. By 1963 an interim capability had been achieved, but not until recently did a high-confidence system exist. The first two operational vehicles of the satellite early-warning network were launched on May 5, 1971, and were placed in near-synchronous orbits over the Indian Ocean at about 65 degrees east longitude. Their orbits have a 10-degree inclination to provide more northern coverage than a truly synchronous orbit would allow. A slight ellipticity of orbit increases the time they spend over the Northern Hemisphere. Two satellites are required to provide continuous coverage. Although the primary mission of these satellites and others deployed at different longitudes is to provide early warning of an ICBM attack, they are of interest here because they also provide a capability to monitor Russian missile tests.

Both the OTH radars and the early-warning satellites can be used to help verify the testing restrictions included in d



then transmitted back to the earth, where they are converted to black-and-white images. The latter in turn can be used to make color composites by registering the images sequentially on color film through the appropriate filters. These particular photographs are shown here only to help illustrate the multispectral technique; their ground resolution (roughly 200 feet from an altitude of 569 miles) in no way compares with the ground resolution obtainable with the current generation of close-look reconnaissance satellites (roughly one foot from an altitude of 100 miles). The region viewed in the photographs includes Vandenberg Air Force Base (*arrow*).

the SALT I ABM treaty. Beyond that they can be used in conjunction with land-based line-of-sight radars to indicate when and where missile tests are taking place and to provide some information about the type of test. These systems would therefore be used to monitor any agreement that imposed numerical or geographical restrictions on missile-testing. As long as a missile test is within view of a line-of-sight radar its point of ballistic impact can be determined. Some characteristics of the reentry vehicle, such as its size and shape, can also be determined from radar observation. The precision of such determinations would depend on the detailed characteristics of both the radar and the reentry vehicle.

The most useful observations for monitoring long-range missile tests are those made from ships and aircraft in the region of impact of the reentry vehicle. From terminal radar and photographic observations detailed information about the reentry system can be derived. If one compares the calculated impact point with the observed impact point and knows the local weather conditions at the point of impact, one can make estimates of the mode of reentry through the atmosphere. For example, one can estimate whether the reentry vehicle has been designed to glide past its ballistic-impact point or to fall short of it. From the close-range radar pictures of the reentry vehicle and its observed speed of passage through the atmosphere, estimates can be made of its weight and ballistic coefficient (a constant that represents the efficiency of the reentry vehicle in overcoming air resistance). If a powered terminal maneuver is attempted, observing radar and infrared sensors should be able to detect it. Multiple-warhead tests can be easily detected near the impact site, if they are not detected earlier by other techniques.

With all this information readily available to observers near the impact area, it is not surprising that such observations have been made for many years. Since 1961 U.S. ships have monitored Russian missile shots into the Pacific. These observations not only have allowed the intelligence community to keep abreast of qualitative improvements in Russian missile technology but also have provided data on the radar "signatures" of Russian missiles, supplying an important input for the design of ABM radars. The sophistication of current capabilities was demonstrated on April 23, 1970, when a set of photographs of a Russian multiple-warhead test was released to the press by the Department of Defense. It was reported that after shipboard radar had detected the incoming reentry vehicles, aircraft and vessels were dispatched to make the photographs.

The verification of any qualitative restrictions on ICBM's that might emerge from SALT II would rely chiefly on these land-based line-of-sight radars, OTH systems, early-warning satellites and shipboard sensors. An examination of their possibilities and limitations is important to any realistic assessment of what may be possible and desirable in SALT II.

One approach that has been suggested in order to inhibit qualitative improvements in offensive missiles is to restrict the number of missile tests that each side is allowed each year. It is argued that if the restricted number were small enough, the competition between the testing of existing hardware and the development of new hardware would eliminate most or all new developments. Putting aside the questions of uncertainties and asymmetries to which such a limitation might give rise, one can say with high confidence that the limitation itself could be verified. Counting missile firings is just what the early-warning satellites, OTH radar and fixed land-based radars do best. If the purpose of such an agreement is to restrain qualitative improvements, however, a useful extension would be to insist that all sanctioned tests be conducted along designated flight corridors and at preannounced times. Such an extension would

RECENT EXAMPLE of high-resolution space photography is provided by this photograph of a strip of the lunar surface made from the *Apollo 15* command ship after the lunar landing module containing the two Apollo astronauts had already landed on the moon. A panoramic camera with a focal length of two feet, built by the Itek Corporation, was used to photograph the lunar surface from a

facilitate the task of ensuring that qualitative improvements have not been made in spite of the numerical limitations.

Where qualitative restrictions themselves are concerned the verification problem becomes much more difficult. The far-reaching development restrictions built into the ABM treaty, however, may be cause for encouragement. In general a prohibition against new types of boosters, front-end configurations or reentry vehicles could probably be verified with high confidence, particularly if it were imposed on top of the previously suggested restrictions. Since the various identifying signatures of current systems are either well known or will become so with time, any new hardware would be distinguishable by an unrecognized signature. Such an across-the-board standstill might be very useful and perhaps even feasible if both the U.S. and the U.S.S.R. could reconcile themselves to accepting certain technical asymmetries.

Other less restrictive prohibitions can also be considered. For example, since terminal maneuvering can probably be detected, either directly with infrared and radar sensors or indirectly by comparing the calculated and the observed trajectories, such maneuvering might be prohibitable. A prohibition of this type would automatically rule out terminal guidance, which would serve no purpose without corrective terminal maneuvering.

Another possibility, although it would be more difficult to verify, would be to restrict improvements in accuracy. Although an observer can tell where a reentry vehicle lands and probably can tell whether the vehicle went through deliberate terminal maneuvers or exploited an aerodynamic shape in order to arrive at its target, he cannot tell from his own observations alone where it was supposed to land. Therefore information on missile accuracy must come from second-order inference. One way to estimate the accuracy of the guidance systems, but not the reentry techniques, is to monitor the apparent accuracy of the inertial guidance on space missions. Presumably the technology available for space shots is also available for ICBM's. The U.S. has a very capable worldwide network of radars in the Space Detection and Tracking System, and these systems can be used for making such inferences. Nonetheless, trying to limit accuracy by constraining inertial-guidance systems does not seem to be a fruitful approach. Instead, prohibitions against giving reentry vehicles terminal-maneuvering capabilities and high ballistic coefficients might be used to indirectly inhibit improvements in accuracy.

Another concern for the SALT II agenda is multiple-warhead tests. The presence of several warheads can be detected by terminal radar, shipboard radar and possibly by other systems. Could a multiple-warhead system be tested with only one warhead at a time, however, thereby avoiding detection? The full answer must depend on both the detailed structure of the hardware and the confidence required in such a system. A partial answer can be formulated by recognizing that in a world characterized by mutual deterrence the major concern of each side is that the opponent not achieve a capability that can be used for a preemptive, counterforce attack. It seems unlikely that such a capability could be developed and brought to a status of high reliability without a large number of full-system tests. There is a problem, however, in translating the likelihood that the U.S.S.R. could not create such a high-confidence system while it was constrained by an arms-control treaty into a certainty on the part of the U.S. that such a system has not been developed.

In this connection the distinction between MRV's (multiple reentry vehicles) and MIRV's (multiple independently targetable reentry vehicles) may be quite inconsequential, depending on the nature of the systems. For a small separation of the reentry vehicles, systems with either mechanical separation or independent guidance could be conceived. To distinguish one such system from another would be difficult and of little value. For a wide separation, however, independent guidance would be required in order to prevent a degradation of accuracy. Presumably a system that can produce wide separation could also be programmed for small separation and could perhaps be tested without being recognized for what it was. In general it seems very difficult to impose verifiable restrictions on multiple-warhead systems except by a total prohibition or a freezing of present systems.

This article has been concerned with national technical means of verification because these are the techniques mentioned in and protected under the SALT I accords. The conclusion should not be drawn, however, that such technical means of verification are the only means available. Other national (or unilateral, as opposed to on-site, or cooperative) verification techniques include economic analysis, diplomacy, content analysis of documents and speeches, interviewing travelers and participants in scientific meetings, and espionage. Although the utility of these information sources cannot be denied, they do have the disadvantage of relying on inference. Reconnaissance and surveillance, on the other hand, are dependent primarily on the physical properties of electromagnetic sensors and therefore provide less ambiguous information and broader coverage than other techniques.

This discussion cannot even claim to have exhausted the national technical means of verification. Specifically omitted because of the lack of sufficient unrestricted information have been the ships, the aircraft, the satellites and the

height of about 60 miles. The reproduction of the complete strip photograph shown here is approximately a third of the size of the original film. The original has a ground resolution of about three feet, suggesting an optical resolution for the system of better than 180 lines per millimeter of film. The *Apollo 15* landing site is within the small white square at lower center (*see illustration at bottom*).

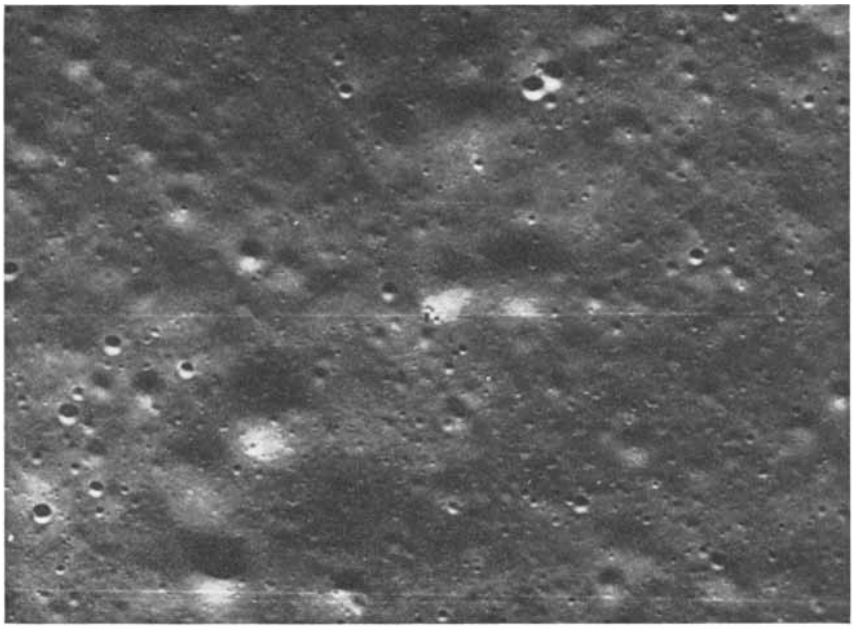
land-based receivers used for electronic and communications intelligence. The information gathered by these devices, such as radar transmission frequencies and the nature of Russian and Chinese communications networks, may be of greater relevance to military planning than to the verification of the SALT agreements. To the extent that communications can be intercepted and decoded, however, the intelligence community may be able to go beyond its assessment of technical capabilities to clues about political intention.

The major conclusion that can be drawn from the analysis presented here is that the U.S. can, with its observation satellites and missile-test-surveillance systems, verify Russian observance of the SALT I ABM treaty and interim agreement with high confidence. Although the focus of this article has been exclusively on the U.S. capability, the Russians have comparable observation-satellite systems that they also can rely on to verify U.S. compliance. The mutual benefits of verifying each other's activities by national technical means have now been formally recognized in the noninterference and nonconcealment stipulations of the current agreements.

The conclusion of verifiability is not dependent, however, on Russian cooperation in nonconcealment. Although it is much easier to monitor a cooperative target nation, the possibility of cheating on an arms-control agreement must not be ruled out. In fact, the purpose of reconnaissance systems is precisely to detect or deter such cheating.

The next major arms-control negotiations, SALT II, are now under way. There is a widespread hope that in this round of talks the interim SALT I agreement can be transformed into permanent numerical limitations on offensive systems. The verification of such an agreement presents the same issues as SALT I has and could also be accomplished with high confidence by observation satellites and missile-surveillance systems. Even a prohibition of land-mobile systems could probably be verified, although a nonzero numerical limitation might be very difficult to monitor because of the intermittent nature of satellite reconnaissance.

The hope has also been expressed that SALT II may be able to restrain the qualitative arms race in offensive missiles. In this connection there may be a rather attractive package of verifiable restrictions, including a blanket prohibition against the testing of new ICBM boosters and new reentry systems. Since each type of booster or reentry system has a set of unique signatures, the introduction of new hardware could be detected. Such a prohibition would best be accompanied by a numerical limit on the number of tests allowed each year and by a requirement that all allowed tests be both preannounced and along prearranged flight corridors. The first limitation would reduce the opportunities for cheating and increase the incentives for compliance, since the testing of new hardware could only be done if the opportunity to test existing hardware were given up. The second restriction would make the task of verifying the prohibition against new hardware easier. Both restrictions could themselves be verified, since with current surveillance systems no unauthorized test could be conducted clandestinely.



BLOWUP of the *Apollo 15* landing site made from the panoramic strip shown at the top of these two pages clearly reveals the lunar landing module (*dead center*). The resolution of this reproduction, made from third-generation negatives, is considerably poorer, of course, than that of the original, in which the individual experiments set out by the astronauts can be distinguished. In other photographs made with this system the tracks of the "lunar rover" used by the astronauts are also visible. Even so, the intrinsic optical resolution of the system is still much poorer than that of the most advanced satellite-reconnaissance cameras, which are probably close to achieving resolutions that are limited only by atmospheric effects.

The Origins of Nerve-Cell Specificity

Individual neurons differ from one another more than cells of other tissues do. Each young neuron apparently obtains a specific "address," which it uses in establishing its associations with other nerve cells

by Marcus Jacobson and R. Kevin Hunt

The development of the brain in the embryo is a tightly orchestrated performance, with the score written in the genes and with a highly predictable and invariant sequence of movements. The neurons, or nerve cells, arise and are sent forth to play their different parts in a regular spatial and temporal order. The young neurons spin out fibers—the axons and dendrites—that extend in precisely defined directions to make contact with other nerve cells. Some of these contacts develop into stable synapses, the points of communication between neurons. In this way the main neuronal circuits achieve an architecture that is breathtaking in its complexity but frugal in its variability: the locations of nerve cells, the trajectories of nerve fibers and the spatial arrays of synaptic connections are invariant in all individuals of the same species. This invariance is termed neuronal specificity.

The orderly development of brain circuits depends on the differentiation of cells with specific properties at particular locations and on the formation of spatially ordered connections between selected cells. The essence of such spatial patterning of nerve-cell differentiation is that the individual developing cell acquires some property that is related to its position in the multicellular array. Let us consider a specific example: the retinal ganglion cells of the frog's eye. These cells differ from other retinal neurons in that they occupy a discrete layer of the retina and have long axons, which together constitute the optic nerve connecting the retina to the optic tectum and other visual centers of the brain.

Individual ganglion cells connect selectively with neurons in the optic tectum of the brain in such a way that the topography of the ganglion cells' positions in the retina is mirrored in the topography of their connections in the optic tectum. One obvious question is therefore how the ganglion cells become different from other types of neuron and, on a finer scale of differentiation, different from one another. A second question is how the axons of these cells are able to select pathways from the eye to the visual centers in the brain. A third question is how the axons, the individual optic-nerve fibers, terminate and form synaptic connections on particular kinds of nerve cell at precisely specified positions in the visual centers. Although different mechanisms operate at each stage, they all result in the development of spatially ordered arrays of nerve-cell bodies, axons and synaptic connections. That is, the nerve cell exhibits its position-dependence at all stages of its development.

Most of these general concepts derive from the work of Roger Sperry of the California Institute of Technology, whose classic studies on the regeneration of the frog's optic nerve led him some 20 years ago to propose that there must be an extraordinary degree of chemical differentiation among the cells of the nervous system [see "The Growth of Nerve Circuits," by R. W. Sperry; SCIENTIFIC AMERICAN, November, 1959]. "The optic fibers differ from one another in quality according to the particular locus of the retina in which the ganglion cells are located," Sperry wrote. "The retina apparently undergoes a polarized, field-like differentiation during development, which brings about local specification of the ganglion cells and their optic axons. The functional relations established by the optic fibers in the brain centers are patterned in a systematic manner on the basis of this retinal specificity."

In our work at Johns Hopkins University and at the University of Pennsylvania we have been particularly interested in what we call "locus specificity": the position-dependent property of the individual retinal ganglion cell that enables its axon to reach the proper locus in the map of optic-nerve-fiber projection to the tectum. What we want to know is how and when position-dependence is determined in the embryonic retina. Does it involve a unique chemical label on each ganglion cell, as Sperry first postulated? What we do is examine the changes of state through which the retinal-cell population passes as the ganglion cells develop their spatially organized set of locus specificities. We have chosen for this work the African clawed frog *Xenopus laevis.* Its nervous system is built on the same general plan as that of other vertebrates, including man, but in the frog the retina and visual centers are accessible to surgical operations in the embryo, the eyes are easy to graft from one *Xenopus* embryo to another and the grafts are not rejected.

A further advantage of *Xenopus* is the availability of full information on its developmental anatomy. The eye originates as an outpocketing of the brain that reaches the outer layer of the embryo at embryonic stage 22 and then gradually folds inward to form the optic cup [*see illustration on page 29*]. The first ganglion cells are born at stages 29 through 31, to be followed shortly thereafter by the other types of retinal neuron. Cells in a ring at the circumference of the retina continue to divide and give rise to all types of retinal neuron throughout development, so that the oldest neurons lie nearest the center of the retina and the youngest nearest its periphery.

An axon begins to grow from each ganglion cell within a few hours of its birth and reaches the optic tectum about 15 hours later. By recording with a microelectrode directly from the terminals of these optic axons we can map their spatial distribution in the tectum of the tadpole or later in the adult frog. We obtain a retinotectal map by finding the point-to-point correspondence between the position of a visual stimulus presented to the eye and the position of the resulting electrical responses recorded in the tectum. We must emphasize that this is not a map of synaptic connections between retina and tectum, which could only be mapped by recording at a point in the circuit beyond the synapse of retinal axon onto tectal cell. The retinotectal map nevertheless indicates the relative positions of retinal axon terminals in the tectum, and therefore provides information about the position-dependent properties of ganglion cells in different parts of the retina.

Our experimental strategy has been to operate surgically on the embryonic eye and later to map the retinotectal projection after the tadpole metamorphoses into a young frog. The first question was when to operate: at what stage of development was the organism likely to provide meaningful insights into the processes we wanted to explore? Some 30 years ago Leon Stone of Yale University demonstrated the existence of a brief "critical period" in the early embryo during which a change occurs affecting the fate of the entire retinotectal system. He found that if the eye of a newt or a salamander was inverted before the critical embryonic stage, the animal developed normal vision; if the eye was inverted at a later stage, the animal behaved as if it saw the world upside down, and its topsy-turvy visual behavior was never corrected by practice.

About seven years ago one of us (Jacobson) confirmed Stone's observation and showed that the corresponding critical period in *Xenopus* was between embryonic stages 28 and 32. The retinotectal map, assayed with probes that could detect the electrical activity of the

THREE-EYED FROG serves as an assay for determining the extent to which the third eye, which is implanted in the embryo under various circumstances, has attained a complete set of "specificities" governing its relations with cells in the visual centers of the brain. The experimental animal is *Xenopus laevis*, the African clawed frog, in which the eye is easily operated on and grafted in the embryo.

nerve cells, was normal in an adult *Xenopus* if the eye had been turned through 180 degrees at stage 28. When the same operation was performed a few hours later, at stage 30, the retinotectal map was inverted in one axis. The optic fibers occupied their normal positions in the mediolateral (side to side) axis of the tectum, which corresponds to the dorsoventral (up to down) axis of the visual field. The map was inverted, however, in the anteroposterior (front to back) axis: instead of projecting from front to back on the tectum the fibers projected in the reverse order. Inversion of the eye at stage 31 or later resulted in complete inversion of the retinotectal map in both the mediolateral and the anteroposterior axes of the tectum [*see bottom illustration on page 31*]. On the other hand, the gross structure of the eye is determined in the early embryo, so that a rotated eye always remained upside down anatomically, even when the operation was performed before stage 30 and therefore resulted in a normal projection.

What this means is that there is a change in the embryo, during a five- to 10-hour critical period between stages 28 and 32, that alters the response of the retinotectal system to 180-degree rotation of the eye. Our recent work began with a series of control experiments, which showed that the difference between the response at stage 28 and the response at stage 32 was not simply a difference in the decision of surgically manipulated eyes to "reverse" or "not reverse" their retinotectal maps. We removed eyes and then either reimplanted them in a normal orientation or rotated them through a variety of angles from 90 to 270 degrees. We mapped the resulting retinotectal projection after the frog had metamorphosed from a tadpole into an adult. The retinotectal map was always normal, regardless of the angle of eye rotation, when the surgery had been done at stage 28 or earlier. Following surgery at stage 30 or later, the retinotectal map was always rotated by exactly the same angle as the eye had been.

These experiments allow three main conclusions. First, the set of locus specificities that ultimately develops in eyes that are rotated before the critical period is spatially organized in accordance with the postoperative positions of the retinal cells and pays no heed to the cells' preoperative positions. In eyes rotated after the critical period, on the other hand, it is the preoperative positions of the cells that matter.

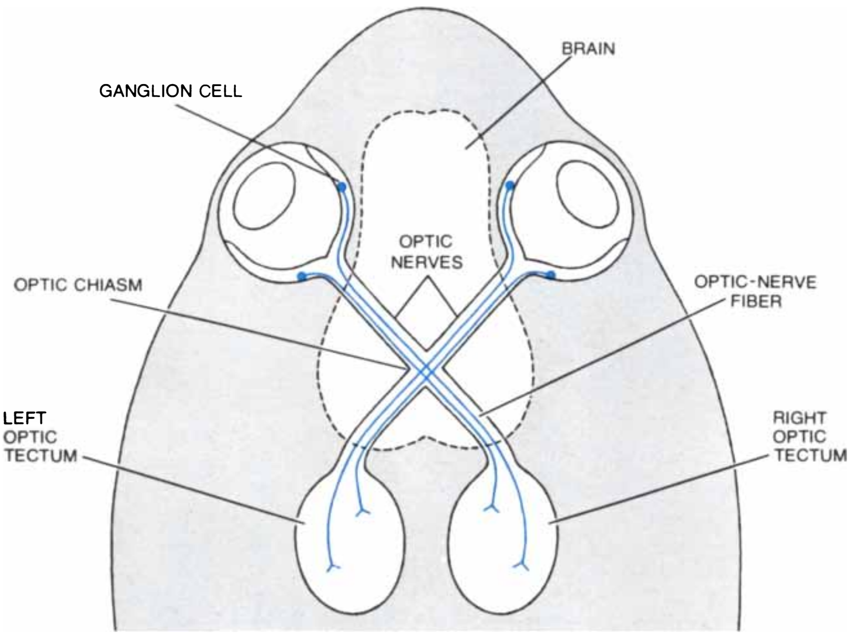
Second, the positional dependence of locus specificity has two axial components, one related to the cell's anteroposterior "address" and the other to the dorsoventral address. The change during the critical period occurs in two steps, each related to one of the axial components. By rotating an eye in the middle of the critical period it is possible to produce ganglion cells with hybrid locus specificities: the anteroposterior component is appropriate to the eye's preoperative position and the dorsoventral component is appropriate to its postoperative position.

Third, even at stage 32, when no retinal axons have yet grown out of the eye and the majority of the ganglion cells have yet to be born, rotation of the eye results in a corresponding rotation of the entire retinotectal map involving all the optic-nerve fibers of the adult eye. Apparently a developmental program has been established in the early embryo that affects the entire set of specificities in the adult retina.

These experiments have served as the cornerstone for our subsequent work. Moreover, they illustrate a peculiar quality of this system: Under many variations of experimental input the system essentially gives rise to a binary output. That is, we change the anatomical orientation of the eye under a variety of experimental conditions and get either a normal map or a rotated map. It is from the contingencies under which the eye was disoriented that we can characterize the intermediate states and changes of state and infer the control mechanisms operating in these transitions.

We can change the orientation of the eye during a transfer between embryos of different stages and correlate the result with the stages of the eye and the host embryo. We can transfer the eye into a tissue culture and change its orientation when we reimplant it in a host embryo, and correlate the result with the results of direct-transplantation experiments. Finally, we can disorient the eye repeatedly during serial transfers from a donor embryo to one or more intermediate hosts and then to a final carrier embryo in which retinotectal connections are allowed to develop; the result of such a serial-transfer experiment can be correlated with the various positions and orientations of the eye during its earlier history.

One such serial-transplantation experiment was designed to show that locus specificities were true position-dependent cellular properties, not simply the manifestation of a correspondence in timing between the birth dates of cells in the retina and the arrival of optic fibers in the tectum. We delayed the arrival of fibers in the tectum by transplanting a stage-32 eye to the side of another embryo and leaving it there for 30 days. During this time many rings of cells were added to the retinal margins and many cells were added to the tectum. When the eye was reimplanted in



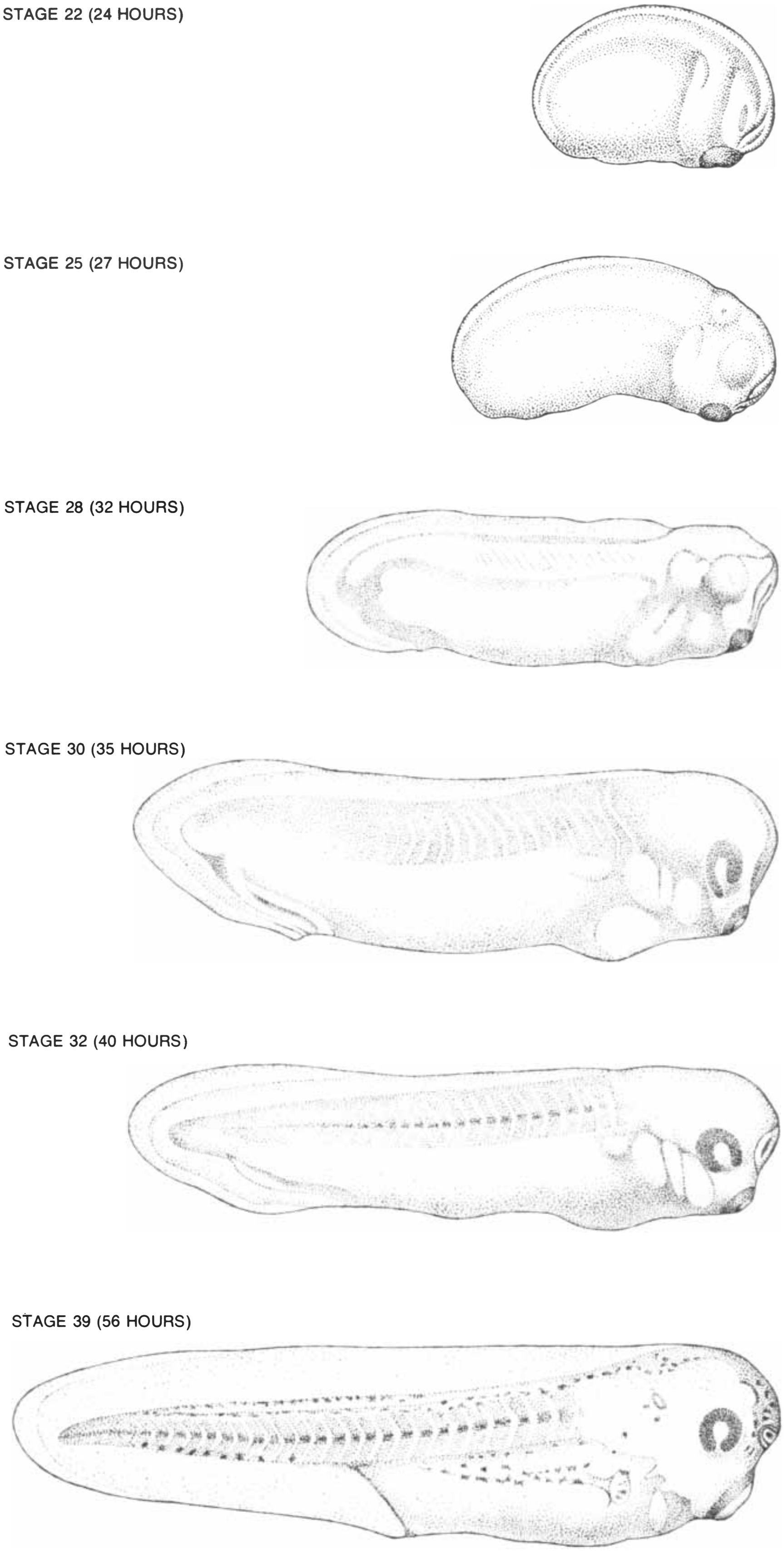
FROG'S VISUAL SYSTEM is diagrammed. Messages from individual ganglion cells in the retina of the eye travel along the cells' axons, the fibers of the optic nerve, to the optic tectum in the brain. The optic nerves cross at the chiasm, so that signals from the left eye's visual field go to the right tectum and signals from the right eye's visual field go to the left.

the socket in its normal anatomical orientation and allowed to form connections with the brain for the first time, it gave rise to a normal retinotectal map [*see top illustration on page 32*].

This result shows that the many cells produced in the retina and tectum over a 30-day period did not become snarled in a "traffic jam" while trying to establish orderly connections. Instead they were able to enter into the spatial relations that produce the retinotectal map, which they normally negotiate in a more gradual manner. Moreover, when the eye was introduced into the socket in an inverted position, the resulting retinotectal projection was orderly but inverted, indicating that the system had completed a program established prior to surgery. These experiments show that locus specificity is a stable cellular property and that just such a stable, position-dependent property must govern the embryonic development of the retinotectal map.

A similar experiment was addressed to the question: On precisely which positions are the locus specificities dependent? Clearly there must be an interaction between the eye and the axial cues of the embryo, because the axes of the eye are aligned with the major axes of the embryo. The question is whether the eye uses the body merely as a source of cues for establishing properly aligned axes of its own (and then specifies individual cellular addresses along its own axes) or whether the individual retinal cells are "instructed" directly by the embryo as to where they lie on the body surface.

To test these alternatives we used embryos that had not yet reached the critical period [*see top illustration on page 32*]. Stage-28 eyes were grafted in various orientations on the sides of stage-28 intermediate-host embryos, left there until well after the critical period was over and then reimplanted in the orbits of final carrier embryos. When the retinotectal projections were mapped after metamorphosis, in every case the set of locus specificities was spatially organized in accordance with the orientation of the eye on the body of the intermediate host. Therefore it is clear that the axial cues are not unique to the tissues that normally adjoin the eye in the head but are also available to an eye on the side of the body. More important, these experiments show that locus specificities are dependent on the position of the ganglion cell within the eye, and that they must be derived from information about the cell's address along the eye's own



DEVELOPMENT of the *Xenopus* embryo is shown at six stages. The drawings are all to the same scale; the embryo is about two millimeters long at stage 22 and six millimeters long at stage 39. For each stage the time since fertilization is given. The embryo is hatched at about stage 30 and begins to eat at stage 42, after which it is considered a tadpole.

axes and not from any absolute value or measure of position on the body surface.

In describing these results and subsequent ones, we have borrowed from the theoretical ideas of Lewis Wolpert of the Middlesex Hospital Medical School in London, who has formulated a theory for the origins of position-dependent differences in embryonic-cell populations. In its very abstractness Wolpert's terminology is compatible with the fact that our experiments provide operational characterizations at best; we must avoid making unwarranted assumptions as to exact mechanisms and schedules by which individual ganglion cells acquire information about their retinal positions and develop corresponding locus specificities. We therefore think of the changes occurring during the critical period from stages 28 to 32 as programming events that commit the retina to permanent reference axes and specify the positional information that all the ganglion cells will use to develop their appropriate locus specificities. We refer to a process or period of "specification," and we call the states of the system before and after the critical period "unspecified" and "specified."

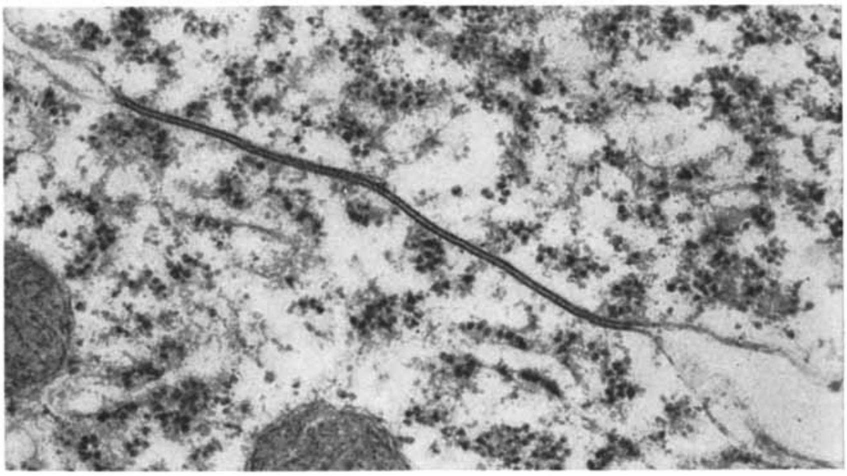
The transition from the unspecified state of the system at stage 28 to the specified state at stages 31 or 32 involves a stable and irreversible change in the retinal-cell population itself [*see bottom illustration on page 32*]. In one experiment a stage-31 eye was "back-grafted" into the eye socket of a stage-28 embryo, which we know contains all the conditions needed to establish properly aligned axes in a rotated or transplanted stage-28 eye. In a second experiment a stage-31 eye was maintained for six to 10 days in a tissue culture before being reimplanted in a stage-39 host socket. In both cases normal retinotectal maps developed from eyes reimplanted in normal orientation, but when the eye was reimplanted in rotated position, the retinotectal map was correspondingly rotated. That is, neither the conditions present in a stage-28 embryo nor total isolation from the embryo was able to reverse or disrupt the specified state that had been attained by a stage-31 eye.

These experiments showed that between stages 28 and 31 the retina undergoes a stable, irreversible change that apparently makes the retinal-cell population refractory to information about later alterations in position. Do the individual cells undergo this change and manifest this stability independently or does the change occur in the retinal-cell population as a whole and require the structural integrity of the eye? In order to answer that question we combined the nasal half of a stage-32 right eye with the temporal half of a left eye at the same stage to form a structurally complete eye. If each embryonic half-eye had given rise to the specificities for the corresponding region of a normal eye, the two halves of the reconstituted eye should have projected to different regions of the tectum. Instead each half projected to the entire tectum, with mirror-symmetrical points in each half-retina projecting to the same place in the tectum [*see illustration on page 33*]. Similar results were obtained after fusion of other combinations of half-eye fragments and even after simple surgical bisection of the stage-32 eye. The original program for spatial deployment of locus specificities, in other words, was not expressed after disruption of the structural integrity of the eye at stage 32, suggesting that the stability of the specified state in the stage-32 eye requires a minimum population of retinal cells and does not extend to the individual retinal cell; it may even require the structural integrity of the entire eye.

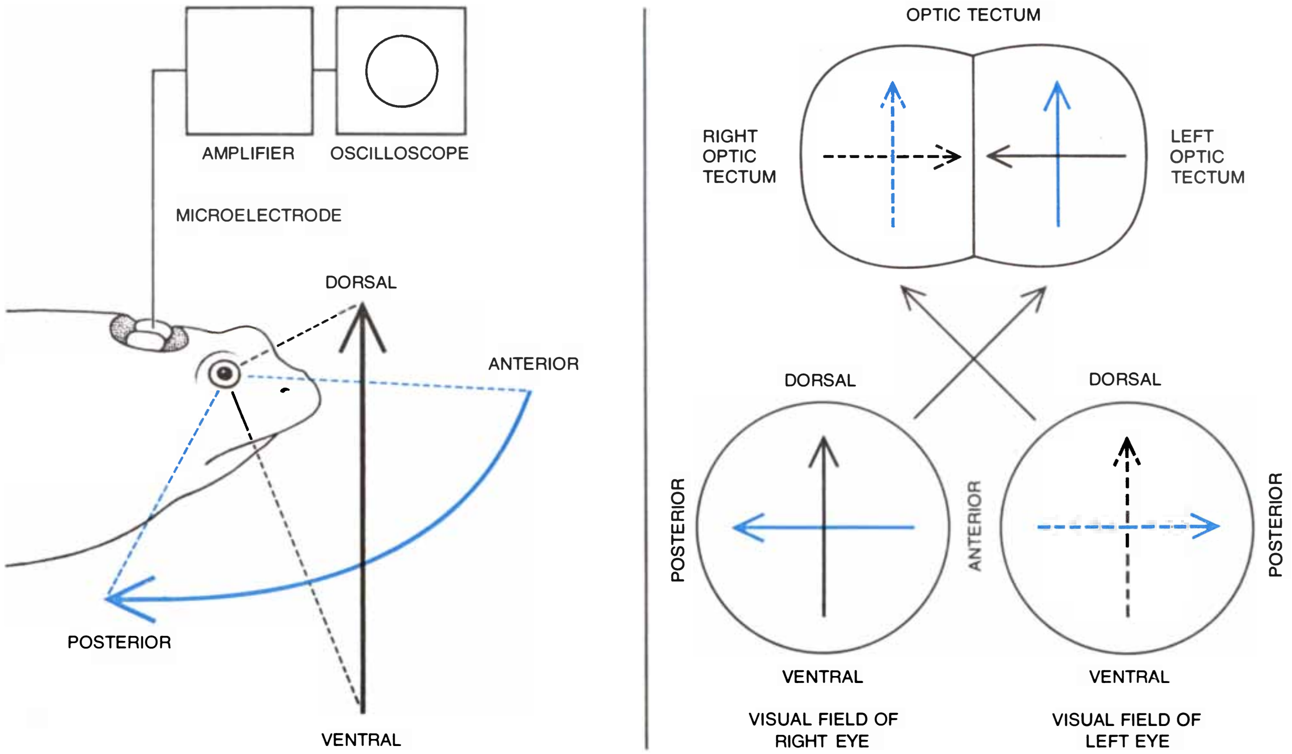
Is the transition from the unspecified to the specified state triggered by mechanisms within the retinal-cell population or elsewhere in the embryo? We transplanted a very young embryonic eye from a stage-23 donor, in rotated position, into a stage-28 host, allowed it to traverse the critical period of that host and then reimplanted it in a stage-32 final carrier embryo. The reimplantation was done, it is important to note, before the eye itself had reached its critical period. Such eyes always developed normal retinotectal projections regardless of their orientation in the final carrier socket, indicating that the set of locus specificities was spatially organized along reference axes acquired from the final carrier embryo—during the eye's own critical period [*see upper illustration on page 34*].

This result shows that the embryo retains its ability to provide axial cues even after it has passed the normal period of specification. Moreover, precocious specification is not precipitated by the eye's contact with an embryo that is passing through the critical stage if the eye itself has not reached that stage. We infer that the transition from the unspecified state is controlled by mechanisms within the retinal-cell population, which are triggered when the cells reach a predetermined stage of their differentiation.

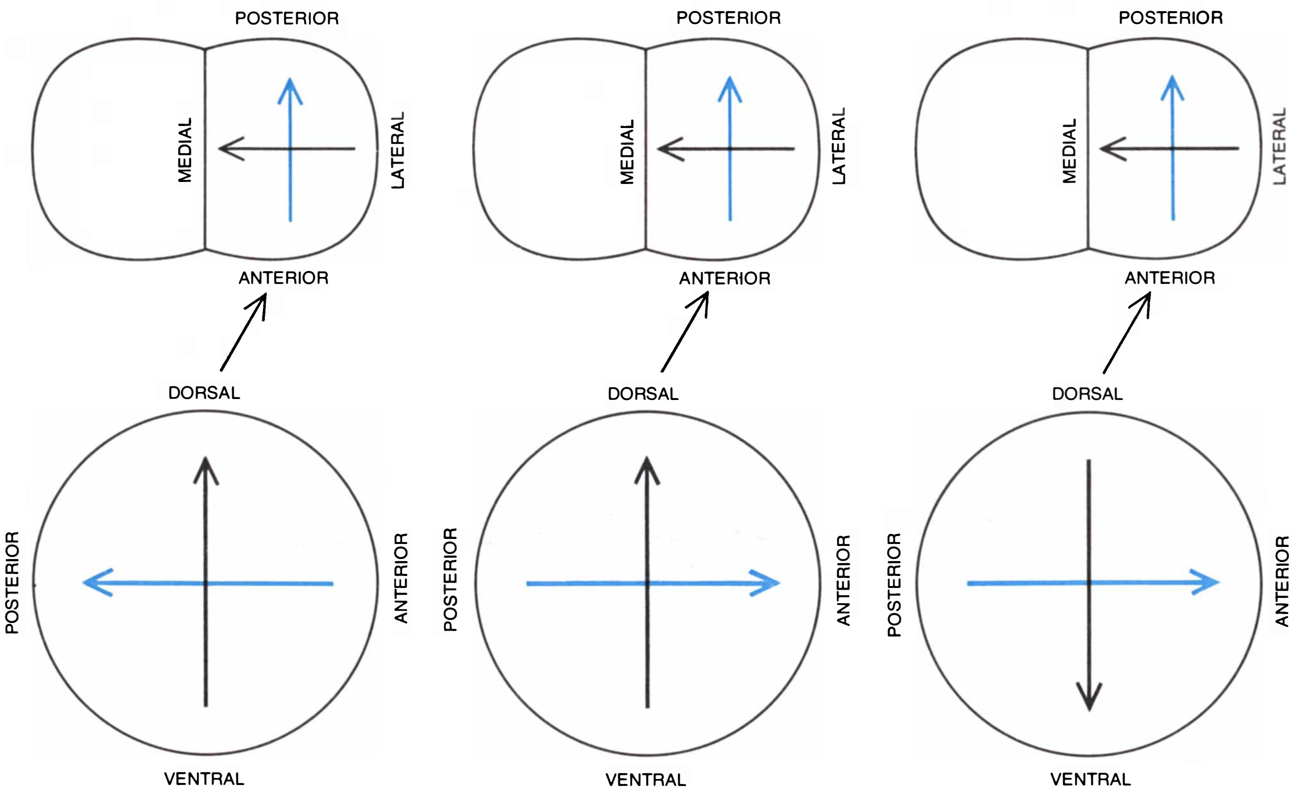
Does the eye possess any information at all about its orientation in the embryo during the unspecified state, before stage 28? Rudimentary eyes from very early embryos (stages 22 to 25) were cultured in glassware for periods of two to four days. Some of these explants attained the form and cellular architecture of eyes that have reached stages 38 or 39; they even spun out extensive optic-nerve fibers in the culture. These particularly healthy and mature explants were then reimplanted in stage-39 carrier embryos. When the retinotectal projections were mapped after metamorphosis, the maps were normal for eyes that had been reimplanted in the final carrier in normal anatomical orientation but were inverted for eyes that had been reimplanted in a rotated orien-



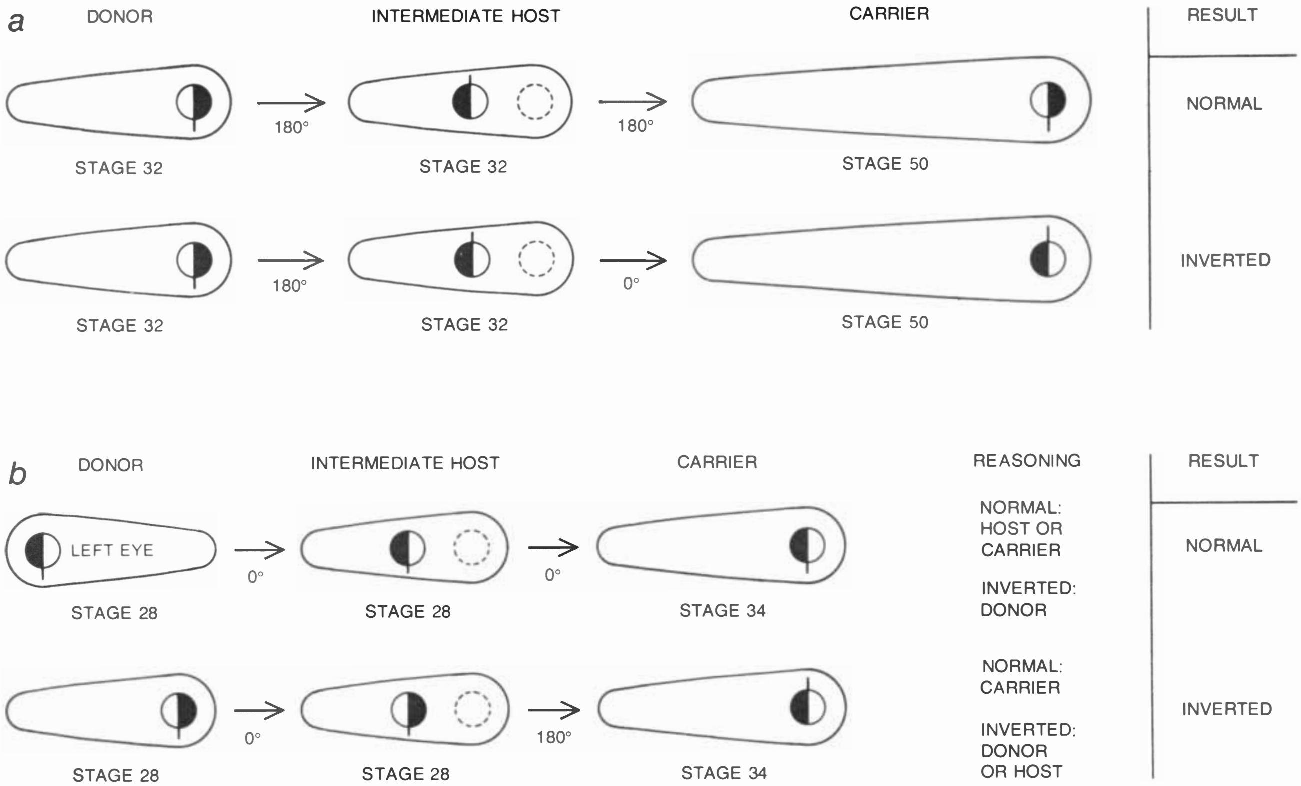
GAP JUNCTION shows as a thick double line in an electron micrograph of two retinal cells made by Marija Duda. At the junction the space between the cells narrows from more than 100 angstrom units to about 30. Such junctions are seen between most retinal cells at stage 28 but are less frequent at stages 29 and 30 and are absent between nondividing cells after stage 31. The junctions apparently serve as channels for intercellular communication.



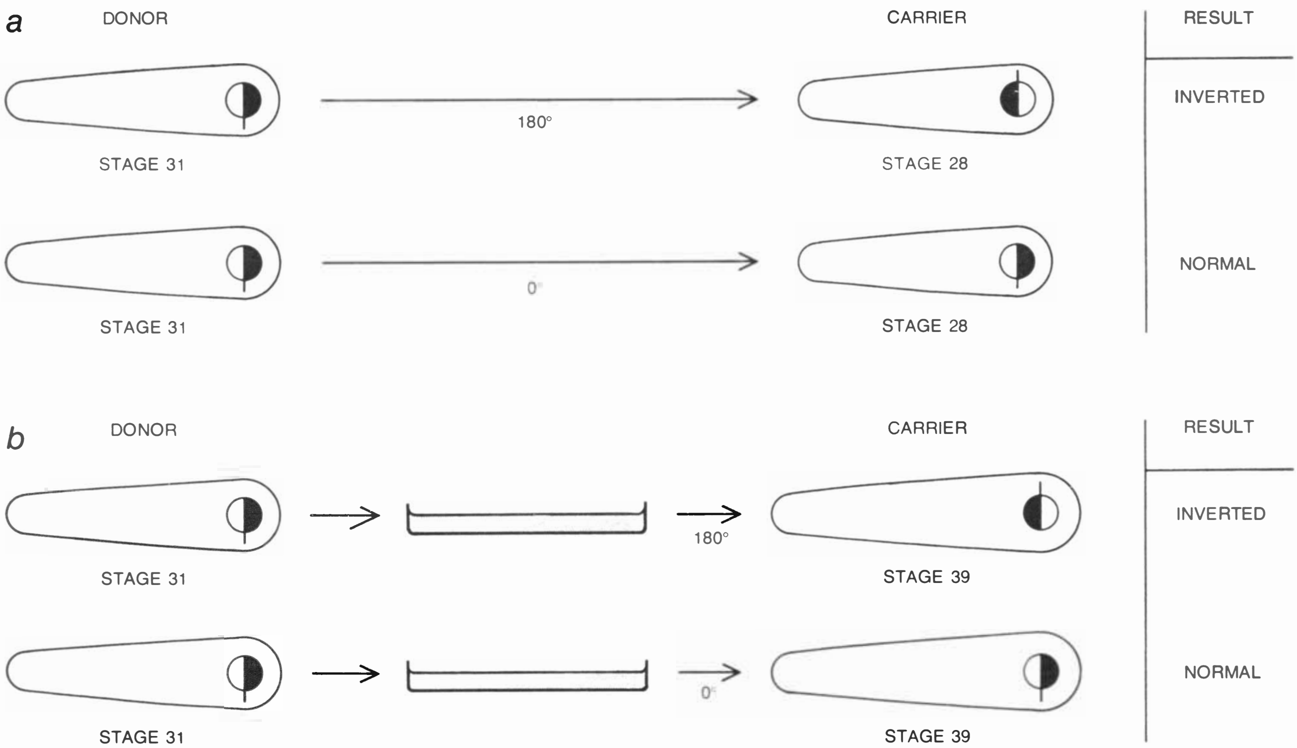
RETINOTECTAL MAP relates the visual field to electrical responses in the tectum. The field is defined by two axes: the dorsoventral, or up-to-down, axis and the anteroposterior, or front-to-back, axis (*left*). These axes correspond to two in the tectum, as shown at right. The anteroposterior axis projects from front to back in the tectum; the dorsoventral axis runs from side to side.



INVERSION OF EYE at various stages of embryonic development has different effects on the retinotectal projection as mapped later in the frog. If the right eye is rotated 180 degrees at stage 28, the retinotectal map is normal (*left*). If the eye is rotated at stage 30, the map is inverted in one axis, the anteroposterior (*middle*). If the eye is rotated at stage 31 or later, the map is completely inverted (*right*). Changes in the projection of the visual field to the tectum are shown here as changes in the visual field, not the tectum, because the experimental procedure involves moving the visual stimulus while maintaining the electrode's position in the tectum.



SERIAL TRANSPLANTS demonstrate the stability (*a*) and the origin (*b*) of positional information. An eye denied tectal connections for 30 days and then reimplanted forms an orderly map that is oriented in accordance with its anatomical orientation (*a*). Eyes that spend the critical period outside the socket also develop orderly maps (*b*). What is the source of their specificity: donor, host or final carrier? The reasoning as to relation between result and source is given. The results show that in both cases the map's orientation is appropriate to the orientation of the eye in the host, which is to say the eye's orientation during the critical period.



CRITICAL PERIOD between stages 28 and 32 involves a stable and irreversible change in the retinal-cell population, as shown by a "back-grafting" experiment to an earlier embryonic stage (*a*) and a tissue-culture experiment in which the eye was maintained in laboratory glassware for six to 10 days (*b*). In both cases the orientation of the map corresponded to the anatomical orientation of the eye in the final carrier. The specified state had been attained by stage 31 and was irreversible: neither encountering a stage-28 embryo (*a*) nor being totally isolated from the embryo in tissue culture (*b*) could disrupt the specified state once it had been attained.

tation [*see bottom illustration on next page*]. Moreover, even when the experiment was repeated with stage-28 hosts, which are known to contain all the conditions necessary for establishing a new set of axes in an unspecified eye, the results were identical.

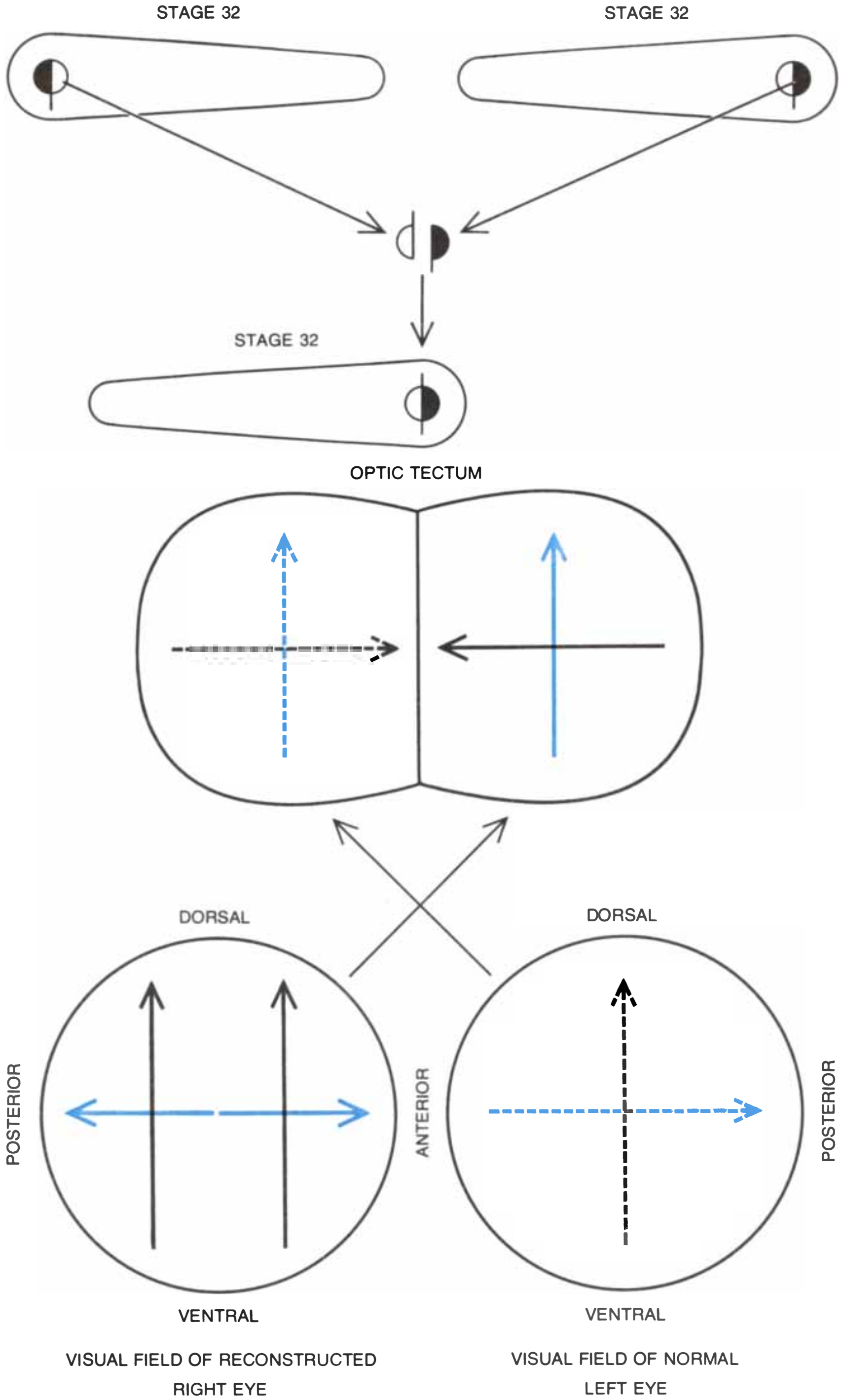
These experiments show that specification can occur in tissue culture in total isolation from the embryo, that the eye as early as stage 22 contains a set of reference axes properly aligned with the axes of the stage-22 embryo, and that specification in a culture results in a set of locus specificities that are spatially organized around these early axes of the eye. The specification process itself—the transition during the critical period from stage 28 to stage 31—is then primarily a change from a reversible state to an irreversible state, and it requires no instructive action on the part of the embryo. The unspecified state must be a stable but reversible one: the axial relationships in an unspecified eye can survive in a culture until the eye is ready for specification, but they can be replaced by new axial relationships following direct transplantation of the eye.

In summary, let us consider the possible changes of state and control mechanisms in the development of a normal retinotectal projection from an eye rotated at stage 25. Initially the rudimentary eye is in a stable but reversible state, having a primitive set of right-angle axes aligned with the anteroposterior and dorsoventral axes of the embryo. An eye rotated at stage 25 interacts with the axial cues of the embryo and the initial state gives way to a second state, which is also stable but reversible. The original set of axes has been replaced by a new set of axes in the eye, which are once again aligned with the major axes of the embryo.

The next transition, triggered when the development program in the retinal-cell population reaches a critical point, "locks in" the anteroposterior axis of the eye and specifies it as the reference axis for the anteroposterior component of positional information for the individual ganglion cells. The third state, then, is characterized by a stable but reversible dorsoventral axis and a stable and irreversible anteroposterior axis. Finally another trigger mechanism is set off, also under retinal control, that locks in the dorsoventral axis of the eye as the reference axis for the dorsoventral component of positional information for all ganglion cells. The fourth state is stable and completely irreversible in the intact eye.

In normal development there may be a mechanism for keeping the developing eye aligned anatomically with the axes of the body, thus making continuous realignment of retinal axes unnecessary. (We have, for example, noted that when the eye is rotated through 90 degrees or less, it often rotates back again to its normal anatomical orientation.) On the other hand, three lines of evidence testify to the rapidity with which the retinal axes can be replaced, and thus to the feasibility of such an axial-replacement mechanism for keeping the retinal and body axes aligned during normal development.

First, we have never found partial inversion of either retinal axis in hundreds



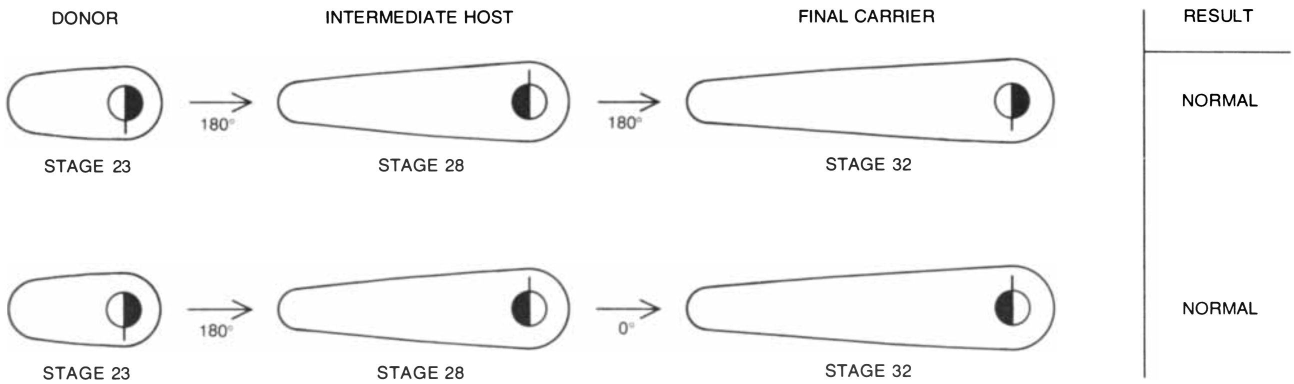
STABILITY OF SPECIFIED STATE was disturbed by an operation in which two half-eyes were combined: a left eye's temporal half, rotated 180 degrees, and a right eye's nasal half in normal orientation. The stage-32 eyes had passed the critical period, but the two halves did not each project to different parts of the tectum. The experiment indicates that the specified state is a property of the whole eye or at least of a minimum population of retinal cells.

of cases of eye rotation, and the retinal and body axes are always perfectly aligned following the inversion of an unspecified eye. Second, less than two hours elapses between stage 30, when eye inversion results in replacement of the dorsoventral retinal axis, and stage 31, by which time the dorsoventral axis is irreversibly specified. Third, axial replacement takes place very quickly in an inverted eye. We rotated an eye for two hours at stage 26 and then transferred it to tissue culture for one or more days before implanting it, in normal orientation, in the orbit of a stage-39 embryo. The resulting retinotectal projection was inverted, which shows that the retinal axes were set up during the two-hour period of eye inversion before the eye was transferred to tissue culture.

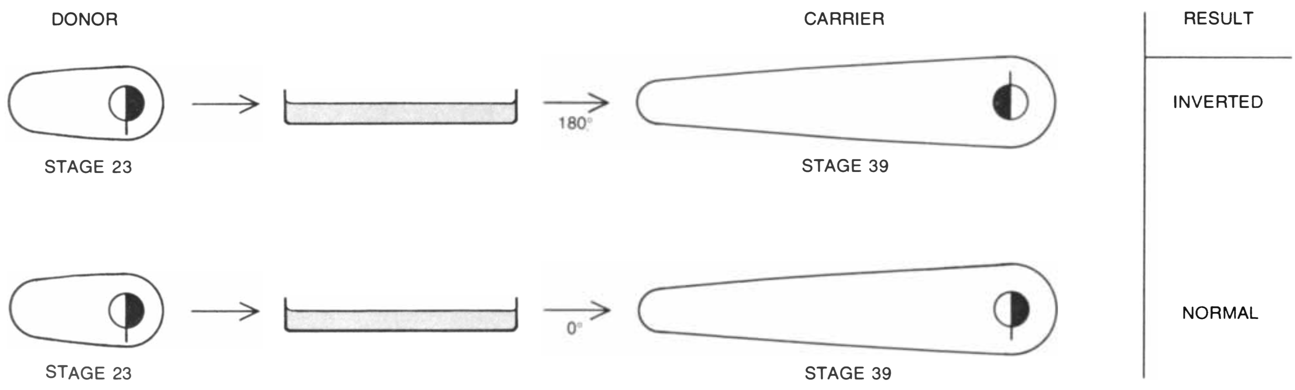
How do ganglion cells that are born in the retina of the tadpole "find out" about the program set up in the early embryo? The transmission of positional information to newborn ganglion cells could be through cell lineage, through intercellular communication or through a combination of the two mechanisms. Each of the maternal cells in the stage-32 retina may be destined to be the ancestor of a line of cells with a preordained set of specificities. This does not appear to be the case, however; the experiment with an eye formed of two halves of a stage-32 retina showed that the fragments do not express the set of specificities predicted from their sites of retinal origin but rather assume a completely different set of specificities. Moreover, interfering with cell production at the margin of the retina (either by disrupting cell division or by killing the majority of maternal cells) does not prevent the development of normally organized retinotectal maps. It is therefore most probable that the newborn ganglion cells at the retinal margin somehow acquire positional information from the older cells in the center of the retina.

The mechanism by which individual ganglion cells process positional information remains to be investigated. Some progress has been made, however, toward understanding the cellular basis of the change that occurs in the retina at stages 29 to 31. Some years ago one of us (Jacobson) showed that this change of state comes about as the cells in the center of the retina stop synthesizing DNA and dividing. The finding focused attention on the events in the last cell-division cycle as the trigger for the change from the unspecified to the specified states.

Recently John Dixon and John Cronly-Dillon of the University of Manchester have made the exciting discovery that certain intercellular junctions disappear from the cells in the center of the retina of *Xenopus* at stages 29 to 31 but persist at the retinal margin. The junctions have since been positively identified as gap junctions, which are generally believed to serve as channels for the transmission of small molecules between neighboring cells [see "Intercellular Communication," by Werner R. Loewenstein; SCIENTIFIC AMERICAN, May, 1970]. We have begun to work with Loewenstein and Birgit Rose of the University of Miami, studying the transfer of small fluorescent-labeled molecules among the retinal cells at various stages of development. Our experiments have yielded encouraging results indicating that cell-to-cell transfer is indeed interrupted at the critical period. It may be that the acquisition of positional information requires intercellular communi-



TRIGGER FOR SPECIFICATION is in the eye itself, as is demonstrated by this experiment. Very young embryonic eyes were serially transplanted to stage-28 and then stage-32 embryos. Specification was not precipitated by implantation in the stage-28 embryo; it awaited the eyes' attainment of their own critical period, which came only after their implantation in the final stage-32 carrier.



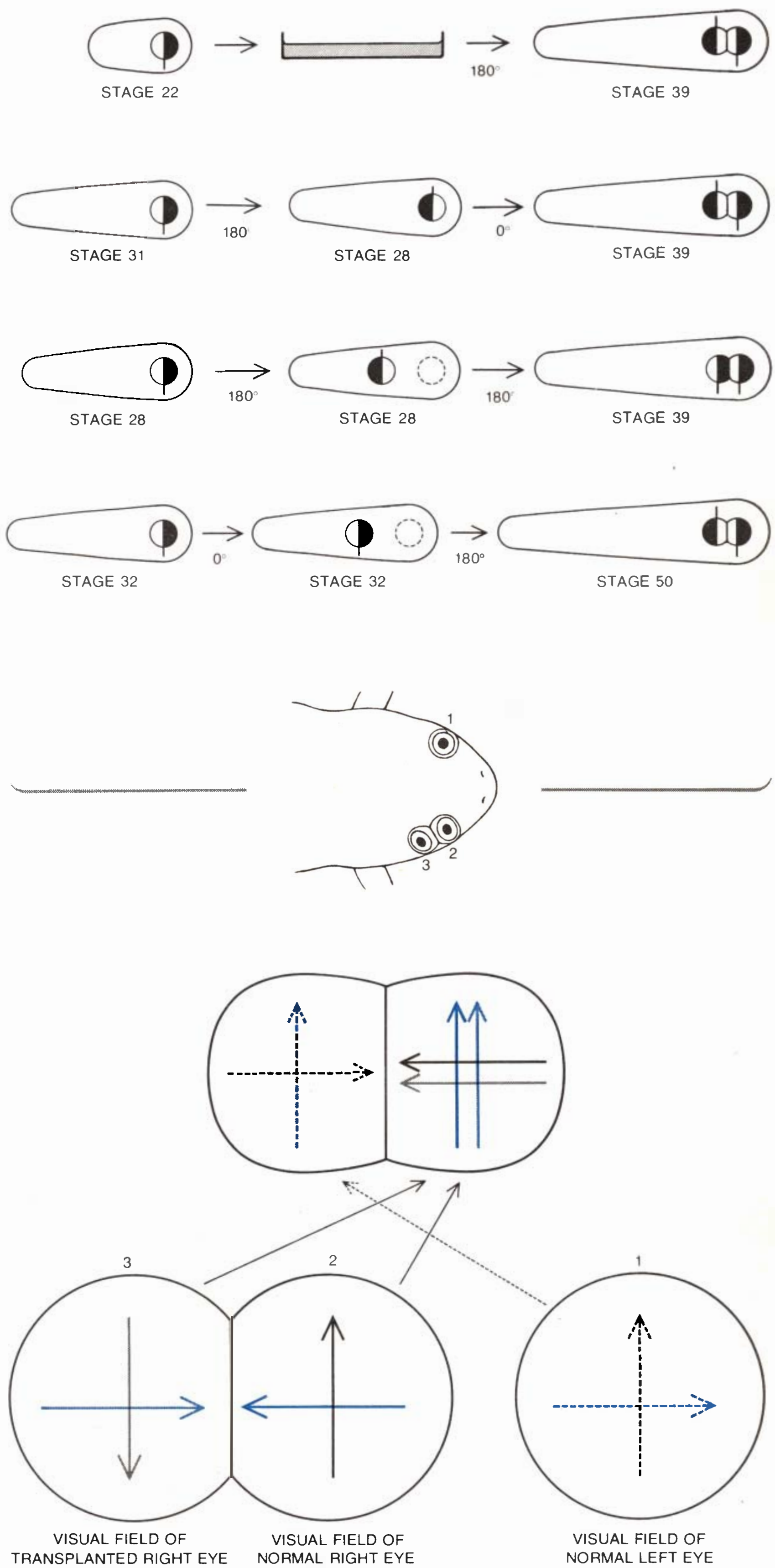
POSITIONAL INFORMATION can be specified in tissue-cultured eyes. Rudimentary eyes were cultured for two to four days and then reimplanted in stage-39 carriers. The results showed that specification had occurred in the culture on the basis of reference axes that had been established in the eyes in the early embryo; specification in the critical period involves a change to an irreversible state.

cation, and cellular uncoupling may be a critical event in the transition from the unspecified to the specified state.

At the beginning of this article we described three stages in the expression of nerve-cell specificity: the acquisition by ganglion cells of locus specificity, the growth of nerve fibers along specific pathways and the selection by the nerve fiber of a target on which to form a synapse. So far we have dealt only with the acquisition of retinal locus specificities, because the phenomenon has a firm observational base from which to make inferences and design other experiments. Much less evidence has been obtained about the selection of axon pathways and the sites of synaptic connections. One reason for this paucity of evidence is that we do not yet know the rules by which retinal locus specificities are expressed during development of the retinotectal circuits. Although the retinotectal map enables us to make limited inferences about some variables, such as the direction of the axes, the map cannot show which specificities are present.

In order to overcome this limitation we have devised an assay that should help us to determine which specificities are attached to the retinal ganglion cells that compose the retinotectal map. In this assay we place a normal eye and an experimentally manipulated eye in the same socket and allow them to project their axons into the same tectum. The projection from the normal eye serves as a standard, and the extent to which the fibers from the two eyes are mingled in the parts of the tectum that are shared gives a measure of the completeness of the set of retinal locus specificities in the operated eye. In all the experiments in which we have either rotated, translocated or cultured an eye, the normal and operated eyes mapped in register to the same tectum, with no tectal region unshared [*see illustration at right*]. This showed that the operated eye contained a full set of locus specificities.

The three-eye assay should give a precise measure of the set of locus specificities in an operated eye, such as a fragment of the eye or an eye reconstituted from two or more fragments, and thus enable us to discover the rules by which retinal locus specificity is expressed. We shall then know whether individual locus specificities can become manifest independently or whether they are only expressed within the context of the entire set of specificities present. Until these assays have been completed the rules that operate in selective neuronal growth and specific neuronal connectivity will remain an enigma.



THIRD-EYE EXPERIMENT tests the completeness of the set of specificities in the third eye, which is implanted in the same socket as a normal eye and projects to the same tectum. After each of the procedures illustrated (*top*) the operated eye mapped in register with the normal one, indicating that it must have contained a normal set of locus specificities.

FISHES WITH WARM BODIES

Not all fishes are cold-blooded. In some fast-swimming species of tuna and mackerel shark a "wonderful net" of arteries and veins conserves the heat of metabolism to increase the power of the swimming muscles

by Francis G. Carey

Fishes are regarded as cold-blooded animals, that is, animals whose body temperature is the same as the temperature of their surroundings. The fact is that not all fishes are cold-blooded. The first man to write about a warm fish was a British physician named John Davy. He was voyaging in the Tropics in 1835, and the ship's company was supplementing its rations by fishing. Davy was intrigued by the copious blood and mammal-like red flesh of one of the fishes brought aboard, a species of tuna known as the skipjack. He took the temperature of several skipjack and discovered that they were warmer than the water they swam in by 10 degrees Celsius (18 degrees Fahrenheit). In the years since Davy's observation most other species of tuna have proved to be warm-blooded and so has one family of sharks: the mackerel sharks. At the Woods Hole Oceanographic Institution my colleagues and I have long been fascinated by these warm-blooded fishes.

What is it that prevents most fishes from maintaining a body temperature that is more than negligibly higher than the temperature of the surrounding water? It is not, as might be supposed, the loss of surface heat that comes from being immersed in a cool medium. Rather it is the heat loss implicit in the fishes' mode of respiration. All land animals and some aquatic ones obtain the oxygen required to support their metabolism by breathing air, which is rich in oxygen and has a low capacity for absorbing heat. Water, from which fish must obtain their oxygen, contains scarcely 2.5 percent of the oxygen found in an equal volume of air but has 3,000 times the capacity of air for absorbing heat. As a result the heat lost in extracting oxygen from water can be 100,000 times greater than the loss in extracting the same amount of oxygen from air.

The oxygen in the water flowing through the gills of a fish diffuses into venous blood pumped to the gills by the heart. The blood may reach the gills at a somewhat elevated temperature as a result of having accumulated metabolic heat. Heat, however, diffuses 10 times faster than oxygen molecules do, so that by the time the blood in the gills is saturated with oxygen its temperature has dropped to the temperature of the water.

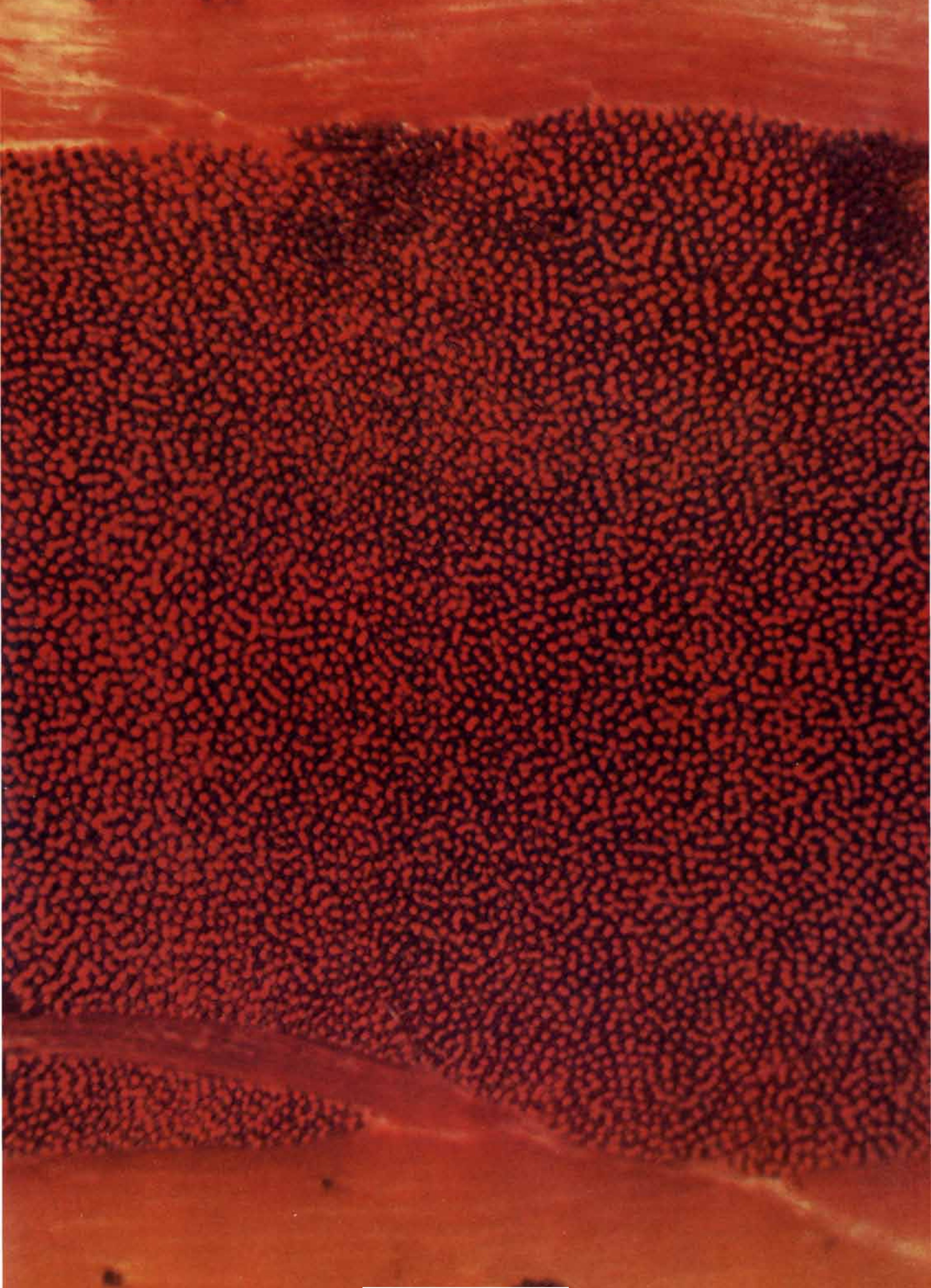
The oxygenated blood now flows back through the fish's circulatory system, and the oxygen is utilized in the body tissues. It is this process that generates the metabolic heat; the rise in temperature is proportional to the amount of oxygen extracted from the blood. In any given volume of oxygenated blood the volume of extractable oxygen is less than 20 percent. Taking into account the usual caloric yield of a fish's foodstuffs, that limited supply of oxygen will support only enough metabolism to raise the temperature of the blood one degree C. Moreover, even that slight temperature increase is lost to the water when the blood next passes through the gills, so that no heat accumulates to keep the fish warmer than the surrounding water.

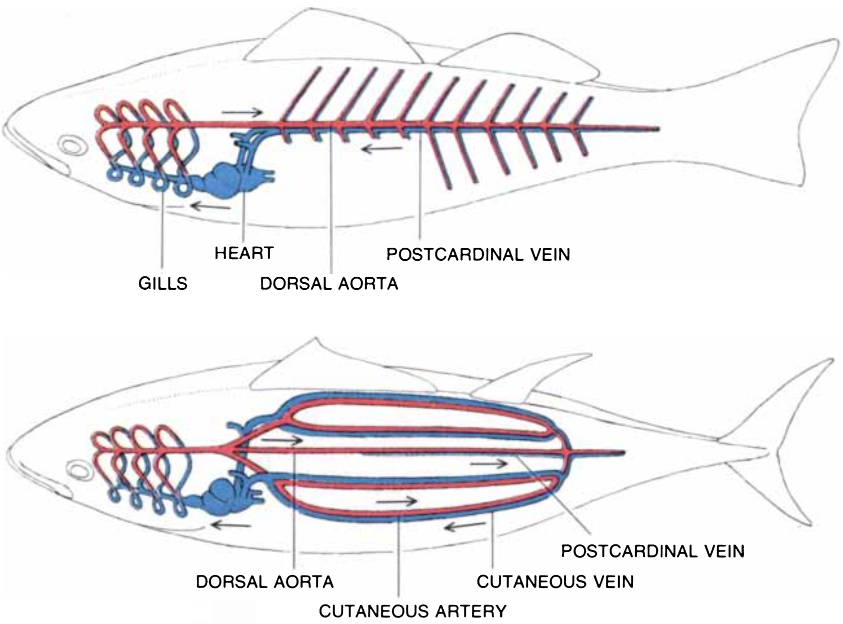
Changes in the fish's rate of metabolism do not affect the cycle, because any exercise or other activity that produces more metabolic heat also demands more oxygen. An increased demand of oxygen means a greater loss of heat through the gills, so that whether the fish is at rest or exercising violently it will always be cold. Indeed, for a fish to raise its body temperature by metabolic processes alone would seem to be an impossibility.

Tunas and mackerel sharks manage to stay warm in apparent defiance of nature because they possess a special structure in their circulatory system. It is the *rete mirabile*, or "wonderful net," a tissue composed of closely intermingled veins and arteries that was first noted in vertebrates by students of anatomy in the 19th century. The French naturalist Georges Cuvier observed the presence of the *rete* in the tuna in 1831, the last year of his life. Four years later, the same year that Davy measured the temperature of the skipjack tuna, the German anatomist Johannes Müller formally described the structures in the viscera of the bluefin.

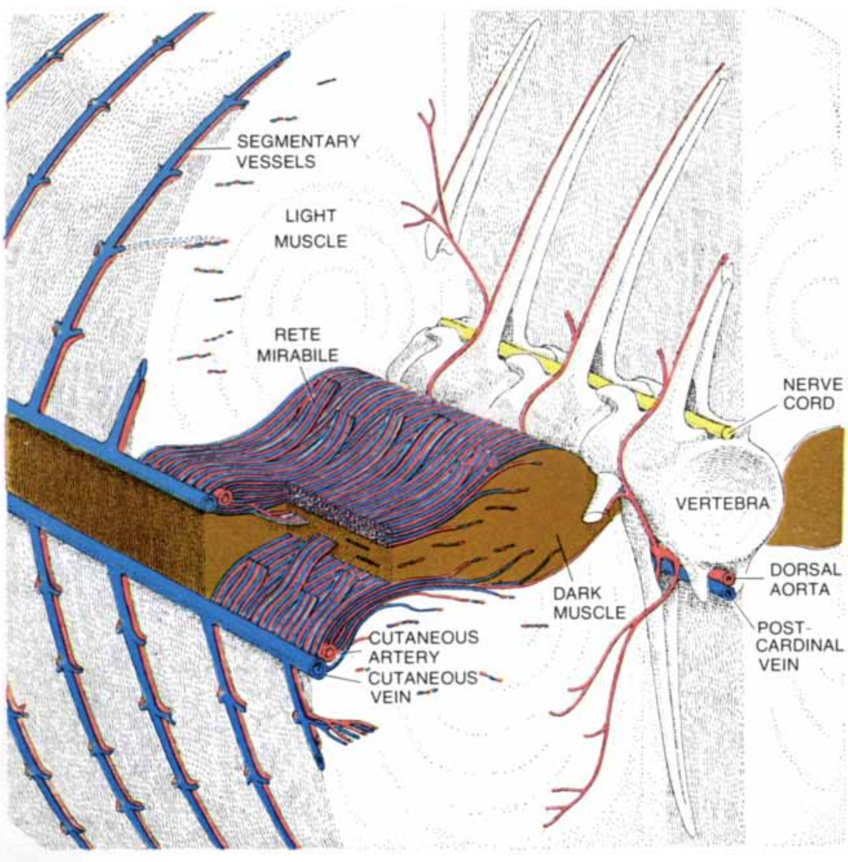
The *rete* provides a thermal barrier against the loss of metabolic heat. The mass of fine veins and arteries permits the free flow of blood for transport of oxygen and other molecules, but at the same time it short-circuits the flow of heat from the body tissues to the gills and shunts the accumulated heat back to the tissues. In the *rete* an artery, carrying cool, oxygen-rich blood from the gills to the body tissues, branches to form a mass of small vessels. The arterial network runs parallel to and is intermingled with a similar mass of fine veins carrying warm, oxygen-depleted blood from the tissues back to the gills. The system of closely associated arteries and

RETE MIRABILE, or "wonderful net," of the bluefin tuna is seen in cross section. The *rete* is a system of parallel small arteries and veins that supplies and drains the band of dark-colored muscle that is used for sustained swimming effort. The system constitutes a countercurrent heat exchanger: venous blood warmed by metabolism gives up its heat to cold, newly oxygenated arterial blood fresh from the fish's gills. The effect is to increase the temperature and thus the power of the muscle. In this bluefin *rete* sample the vessels have been visualized by the injection of latex: red latex in the arteries and blue latex in the veins.





VASCULAR SYSTEMS of cold-blooded (*top*) and warm (*bottom*) fishes are different. The main vessels in most fishes, the cold-blooded ones, run along the backbone and radiate outward to the small vessels (*not shown*) that supply the muscle. In warm fishes the central vessels are smaller; most of the blood flows through cutaneous vessels under the skin and thence through countercurrent nets (*not shown*) that supply the muscle. This arrangement puts the cold end of the exchanger near the skin and the warm end in warm tissues.



BLOOD SUPPLY TO MUSCLE of the bluefin tuna is primarily through four pairs of cutaneous vessels. A multitude of smaller vessels branch from them, forming the thick vascular slab above and below the dark muscle, the *rete*, from which the dark muscle is supplied. The light muscle is supplied by bands of alternating arteries and veins that slant through the muscle from the segmental vessels, thus serving as two-dimensional heat exchangers. Other, non-heat-exchanging arteries and veins run out from the central vessels.

veins acts as a countercurrent heat exchanger. There are large areas of contact between the two kinds of fine vessel, and heat is readily transferred through the vessels' thin walls. Therefore the warm venous blood is cooled as it flows through the *rete,* so that little or no heat is lost when the venous blood at last reaches the gills. At the same time the cold arterial blood is warmed, so that when it reaches the interior of the fish, it has nearly reached the temperature of the warm body tissues.

A *rete* of this type serves exclusively as a countercurrent heat exchanger. Its vessels, although small, are too large and their walls are too thick to allow any significant diffusion of oxygen molecules from the arterial blood to the venous blood. Other *retia* that do serve as gas exchangers are found in many species of fish [see "'The Wonderful Net,'" by P. F. Scholander; SCIENTIFIC AMERICAN, April, 1957]. The tiny vessels that comprise them are capillaries, and so oxygen molecules readily pass from one to another. Such *retia* are found in the swim bladder and the eye.

A major adaptive advantage of an elevated body temperature is greatly enhanced muscle power. If the difference in temperature between two otherwise equivalent muscles is 10 degrees C., the warmer muscle is able to contract and relax three times more rapidly, with no reduction in the force applied at each contraction; thus it can generate three times as much power. This kind of increase in power is evidently essential in flying animals such as birds and bats; their body temperatures are characteristically high. Even some large insects with a heavy wing loading elevate their body temperature when they fly. For example, cicadas and locusts are often unable to take off until they have warmed up by shivering or by basking in the sun.

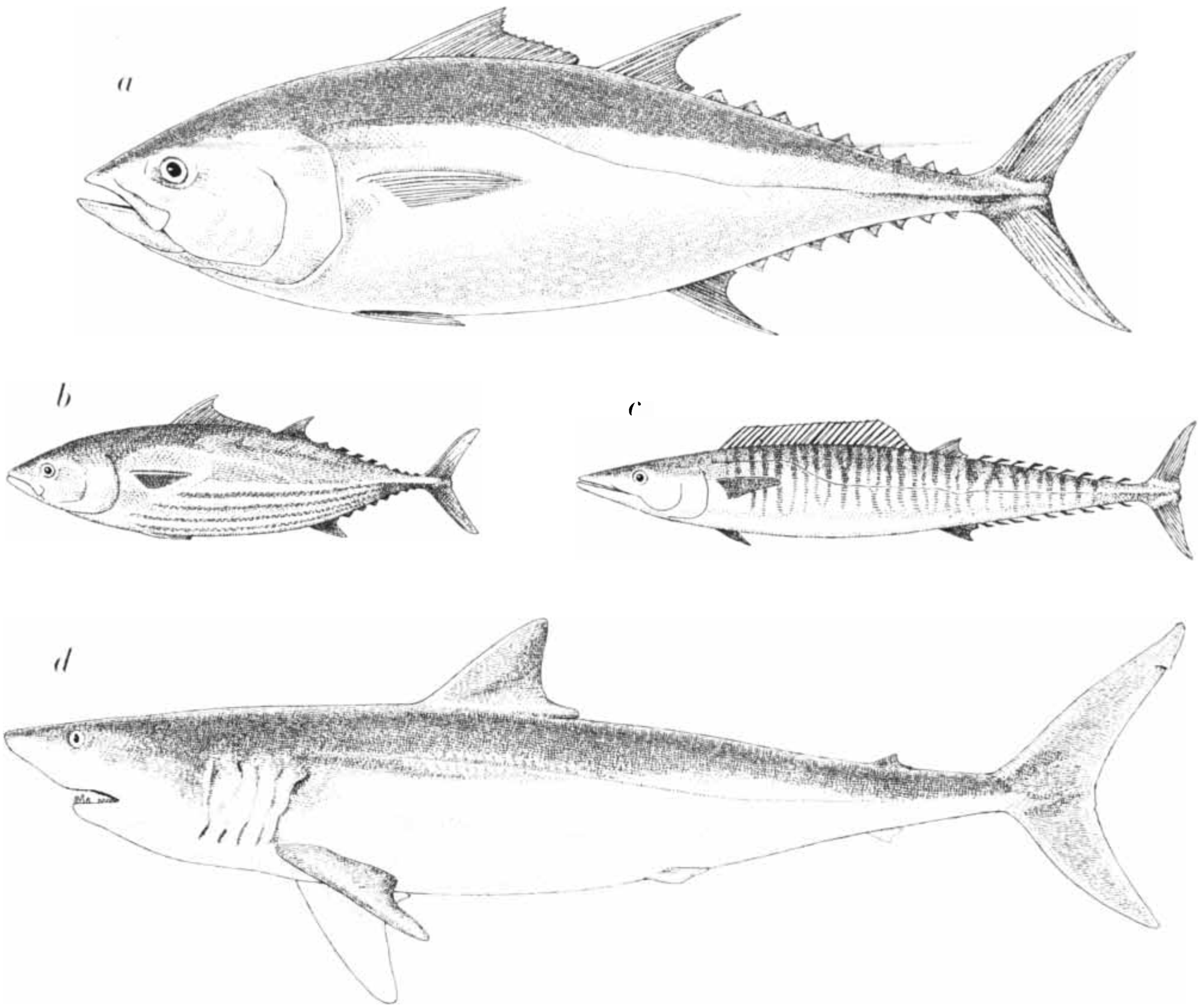
Like flying, high-speed swimming is a demanding form of locomotion. Water is a dense and viscous medium, and to move through it rapidly requires not only fine streamlining but also substantial muscle power. The high body temperature of tunas and mackerel sharks apparently helps to provide the extra power needed for high-speed swimming: the most important of the vascular heat exchangers are those that serve the dark-colored swimming muscles.

Vertebrate muscle is generally a mixture of dark-colored fibers and light-colored ones. The dark fibers contain high concentrations of the oxygen-transporting pigments and enzymes associated with oxidative metabolism; they are also richly supplied with blood. The light-colored fibers usually have a less well-developed blood supply. They can function without oxygen during activity by metabolizing foodstuff through fermentation, recovering later when they are at rest. Muscles that are continuously active, for example the heart, are made up largely or entirely of dark fibers. Muscles that alternate between periods of rest and short intervals of activity contain few, if any, dark fibers.

As anyone who has seen a fish steak knows, the two kinds of muscle fiber are sharply segregated in the axial musculature of fishes. For example, in many tunas a large mass of very dark muscle, well supplied with blood, lies in a broad band between the backbone and the skin along the midplane of the body. This dark muscle can operate continuously; its contractions propel the fish as it swims at ordinary cruising speeds. The blood that supplies the muscle reaches the muscle fibers after passing through a large heat-exchanging *rete*. As a result the muscle is warmer than any other part of the fish's body.

In order to provide for this large heat exchanger the arrangement of the tuna's circulatory system has been drastically altered. Most fishes have a central distribution system for the blood. The main blood supply to the muscle flows outward through a large artery, the dorsal aorta, and returns through the postcardinal vein; both are deep inside the fish just below the backbone. From these central blood vessels segmental arteries and veins radiate out to the periphery.

In the tuna the normal pattern of circulation to the muscle has been almost entirely reversed. The major blood vessels are no longer the central artery and vein. Instead four artery-vein pairs just under the skin, two on each side of the fish, provide the main blood supply [*see top illustration on opposite page*]. From these cutaneous vessels large numbers of tiny blood vessels, each only a tenth of a millimeter in diameter, branch off. The tiny vessels intermingle to form slabs of vascular tissue, the *retia,* that lie close to the upper and lower surfaces of the dark muscle [*see bottom illustration on opposite page*]. These *retia* are the heat exchangers that ensure the warmth of the dark muscle. The blood to the light muscle also comes from the large cutaneous vessels by way of pairs of segmental arteries and veins that run up and down over the surface of the muscle and send numerous branches into it. These branches are in the form of vascular bands: ribbons of alternating arteries and veins that act as heat exchangers.



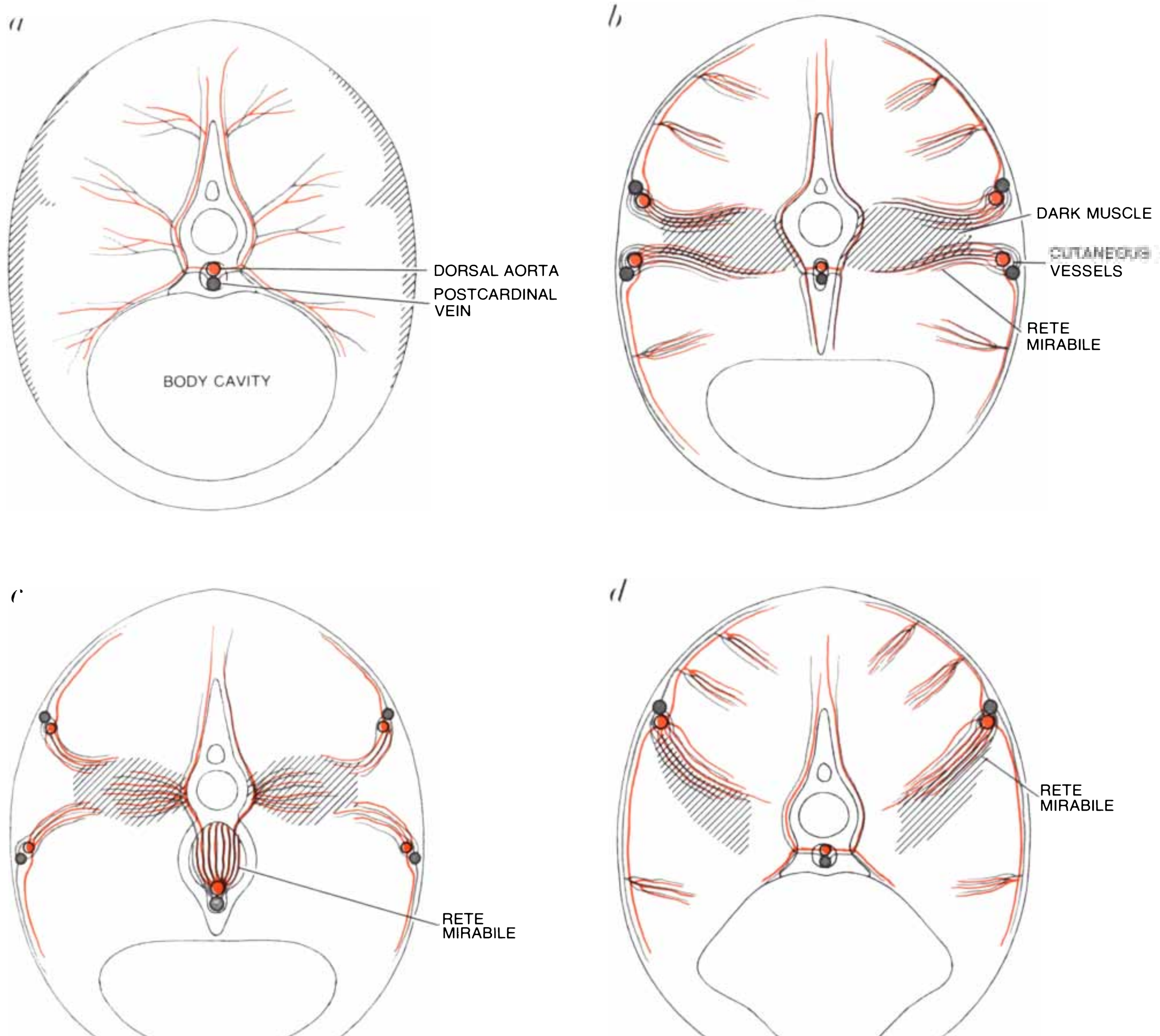
FAST-SWIMMING FISHES with warm bodies have streamlined bodies and heavily muscled tails with crescent-shaped caudal fins. The ones illustrated here are three tunas: the bluefin (*a*), the skipjack (*b*) and the wahoo (*c*), and a mackerel shark, the mako (*d*).

The area of thermal contact between arteries and veins in these two-dimensional structures is smaller than it is in the massive heat exchangers serving the dark muscle, but apparently it is adequate for the typically low rate of blood flow to the light muscle and ensures that much of this tissue too is warm.

One result of the inside-out arrangement of the main blood vessels in a warm fish is that the large arteries and veins that give rise to the *retia* are just under the skin. This puts the cool end of the heat exchanger near the surface of the fish and the warm end deep in the interior, an arrangement that minimizes surface heat loss.

By measuring the temperature of various parts of freshly caught bluefin tuna we have learned that the warmest regions are within the dark muscle [*see illustration on page 42*]. The hot spots are not, as one might expect, in the deepest part of the muscle. The reason is that the deep interior receives some of its blood supply from the dorsal aorta through a system that has no specialized heat exchangers. The highest muscle temperature is found near the center of the dark-muscle band on each side of the fish, but substantial areas of muscle tissue, both dark and light, are considerably warmer than the temperature of the water from which the fish are taken.

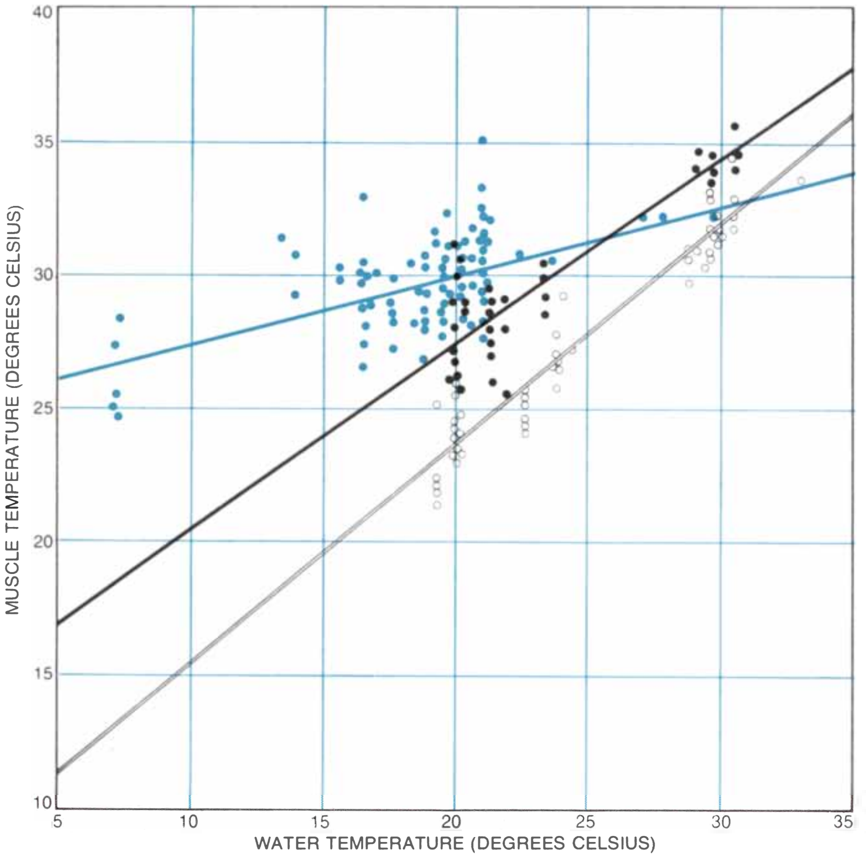
The bigeye tuna and the albacore have circulatory systems that closely resemble the bluefin's. Other tuna species, particularly the skipjack and to a lesser extent the yellowfin, elevate the temperature of the muscle in a different way. Although the pairs of cutaneous arteries and veins are still present, the heat exchangers that arise from them may contain no more than a single layer of fine blood vessels. In these fishes the principal heat exchanger is connected to the central artery-vein pair below the backbone. This central *rete,* a rodlike vascular mass that may be larger in diameter than the backbone itself, extends along the vertebrae in the region above the



MAJOR BLOOD VESSELS of a typical cold-blooded fish (*a*) radiate from the large central vessels. In warm fishes the vascular system is rearranged so that warmed blood is supplied to the dark muscle. In the bluefin (*b*) the *retia* are slabs of vascular tissue above and below the dark muscle. In the skipjack (*c*), on the other hand, the main *rete* is a net of small vessels running vertically in a cavity below the backbone. The mako shark (*d*) has a single cutaneous blood-vessel pair, supplying a single massive *rete*, on each side.

body cavity [*see illustration on opposite page*]. Body-temperature measurements of skipjack attest to the efficiency of the rod-shaped heat exchanger: these little fish are often 10 degrees C. warmer than the water and their warmest muscle is adjacent to the backbone.

It is an interesting coincidence that the mackerel sharks, a group that is evolutionarily far removed from higher bony fishes such as the tunas, should have developed a heat-exchange system that is so much like the bluefin's. The deep interior blood vessels along a mackerel shark's backbone are small. The major blood supply circulates through a single artery-vein pair running just below the skin along each side of the fish. The paired cutaneous blood vessels give rise to a single massive *rete* that warms the shark's dark muscle. In some mackerel sharks, such as the mako shark, the heat exchanger is a solid slab of blood-vessel tissue resembling the bluefin's [*see illustration on opposite page*]. In others, such as the porbeagle shark and the white shark, the *rete* is diffuse, with many bundles of fine blood vessels extending into the dark muscle. In mackerel sharks the body-temperature distribution closely resembles the distribution in the bluefin; most of the sharks' muscle is warmer than the water they swim in, and the warmest parts of all are twin regions within the dark muscle.



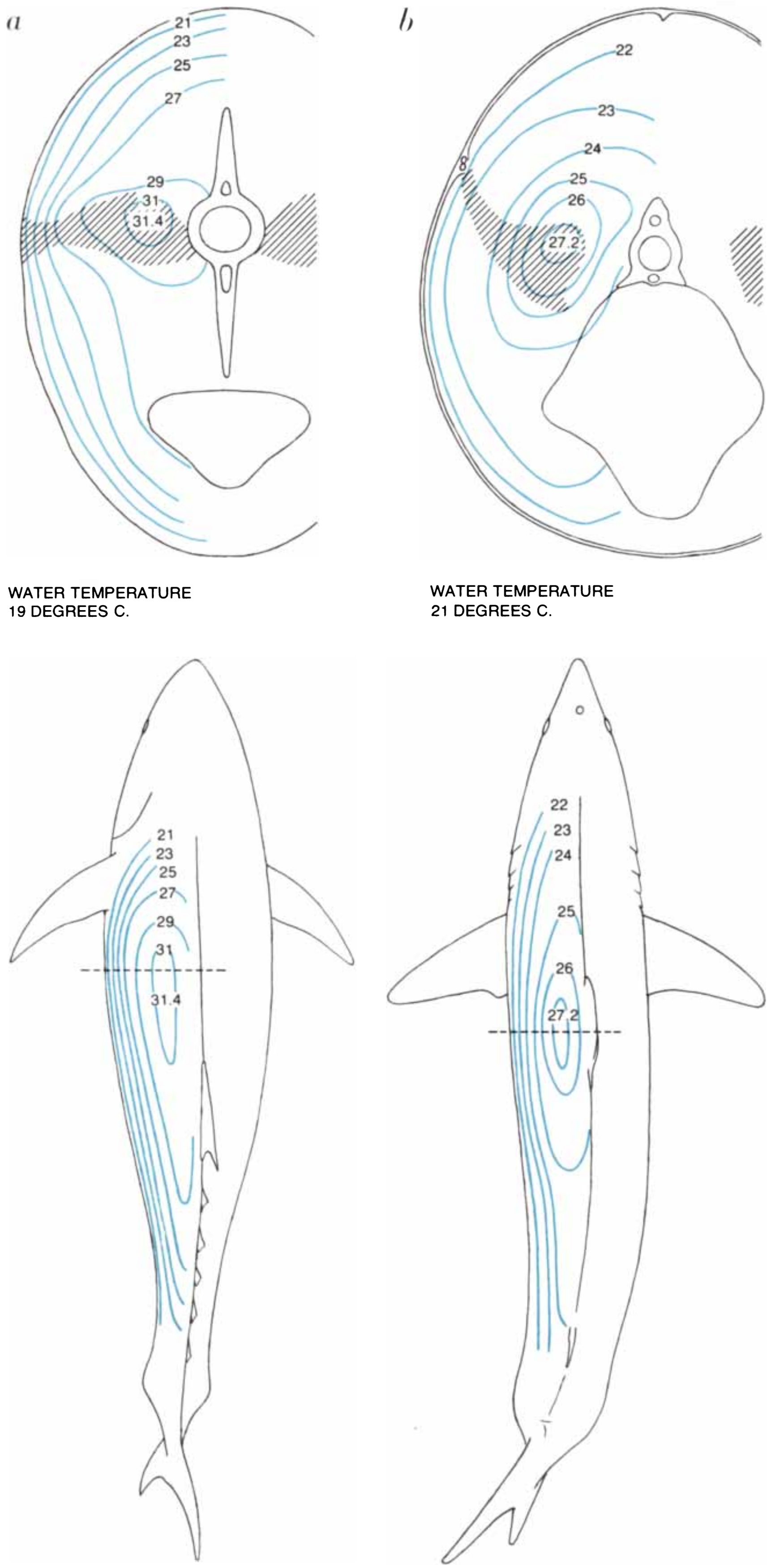
MAXIMUM MUSCLE TEMPERATURES are plotted against water temperature for bluefin (*color*), skipjack (*solid black*) and yellowfin (*open black*). The bluefin regulates its temperature almost independently of water temperature. The skipjack is warmer than the yellowfin but both tend to maintain a fixed temperature difference above water temperature.

Tunas and mackerel sharks possess heat exchangers that are not associated with muscle tissue but serve the organs of the body cavity. In the albacore, bigeye and bluefin tunas dense bundles of intermingled fine arteries and veins are found on the surface of the liver; in a large bluefin the bundles may be five centimeters in diameter. The arrangement in the mackerel sharks is different because the artery that in other fishes carries the main blood supply to the viscera is either very small or nonexistent. Instead the visceral blood supply is provided by arteries that are insignificant in most other fishes. These arteries penetrate into a much enlarged venous space known as the hepatic sinus. Venous blood flows through this space on its way from the viscera back to the heart. The arteries branch and branch again within the hepatic sinus until the space is virtually filled with a spongelike mass of tiny vessels; the cool blood within them is warmed by the bath of venous blood.

Visceral *retia* do not keep the organs of the body cavity warm continuously; the temperature can vary from a level equal to the warmest muscle to one only slightly above the water temperature. The variations may be correlated with the mackerel sharks' and tunas' digestive activity. The tunas in particular have remarkably small body cavities, and the mass of their visceral organs is limited. Raising the temperature of the viscera may speed digestion. Conversely, during times of digestive inactivity the viscera may be allowed to cool.

Since tunas are warm, we wondered whether they could control their temperature at a constant level and thus be independent of water temperature or whether they warmed and cooled as the water temperature changed. Other investigators had noted that yellowfin tuna and skipjack tend to maintain a temperature a fixed number of degrees above that of the water—something that might be accomplished by an unregulated heat exchanger. These two species are found only within a rather narrow range of water temperature, however, and so we turned our attention to the bluefin tuna, which is equally at home in tropical and in subpolar waters. We measured the maximum temperature in the muscle of bluefin taken along the coast from the Bahamas to Newfoundland. Comparison of their muscle temperature with the temperature of the water in which they were captured suggested that they control their temperature quite well. Over a 24-degree C. range of water temperature the average muscle temperature changed only six degrees; bluefin from the 30-degree waters of the Bahamas were only a few degrees warmer than the water, whereas those from seven-degree northern waters were 20 degrees above the water temperature.

There are two ways an animal might achieve such control. One is by means of the swift thermoregulatory adjustment to environmental temperature changes that is characteristic of birds and mammals, which can vary their rate of heat loss in several ways and also generate extra heat by increasing their metabolism. The other is the much slower process of acclimatization, by which cold-blooded animals adjust their activities to changing temperature and which involves cellular and enzyme changes over an extended period of time. In order to discover which of these mechanisms the bluefin were using we decided to try a field experiment in which we would monitor the body tem-



TEMPERATURES were recorded at various points in bluefin tuna (*a*) and mako sharks (*b*). Isotherms show that the warmest parts of the fishes are within the dark muscle on each side of the body. Much of the rest of the muscle is also considerably warmer than the water.

perature of fish as they swam from warm into cold water. If the bluefin were capable of rapid thermoregulation of the kind shown by mammals, their temperature should remain constant during a sudden change in water temperature.

We found that the coast of Nova Scotia was an ideal setting for our project. The Coolen brothers of Fox Point operate pound nets in St. Margaret's Bay, and they frequently catch bluefin tuna in their nets. Moreover, the coastal waters of Nova Scotia are characterized by a marked thermocline (a sharp drop in temperature as the depth increases), so that a free-swimming fish might encounter a wide range of water temperatures.

We used a small harpoon equipped with two thermistors in our first attempts to record bluefin temperatures. One thermistor sensed water temperature and the other muscle temperature; they were connected to an indicator by a 1,000-yard length of field-telephone wire that could unreel as the harpooned fish swam away. We failed on almost every attempt: the fish died or the harpoon was pulled out when the wire was snagged by kelp or by lobster-pot lines. It was at this point that my colleague John Kanwisher suggested that we telemeter the information from the fish.

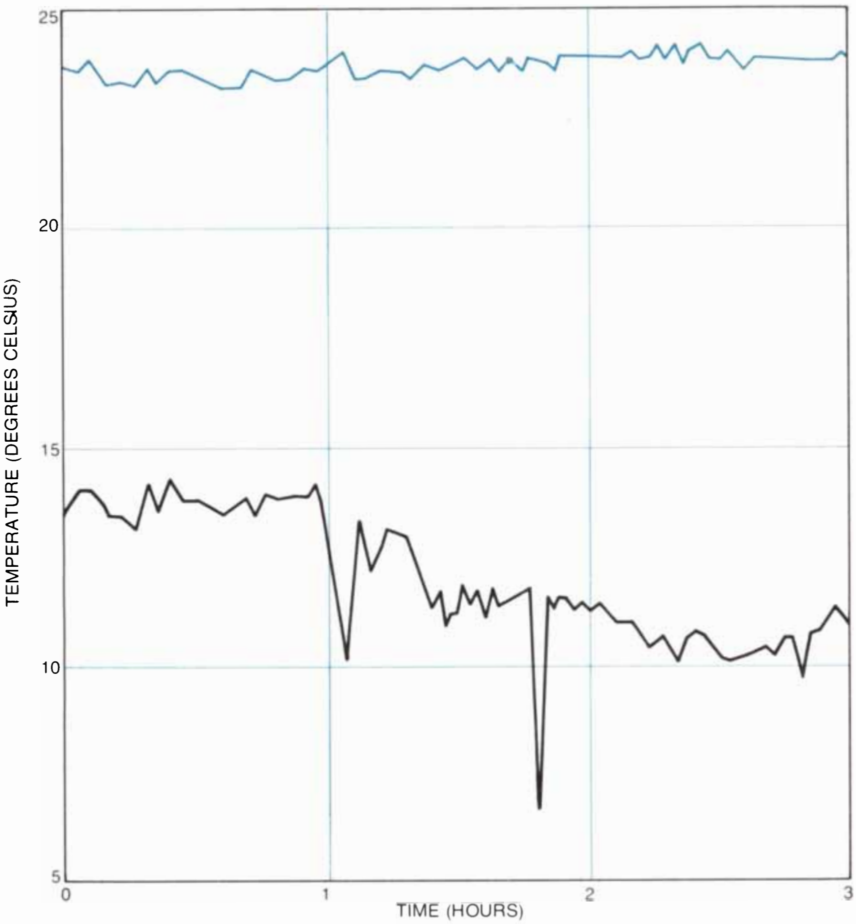
The telemetering devices we selected were battery-powered acoustical transmitters. Their service life was from one day to three days and their range was as much as five miles. The instrument package was mounted on a small harpoon that could be driven through the fish's thick skin with a minimum of injury to the muscle. The thermistor that sensed the muscle temperature was in the tip of the harpoon and the one that sensed the water temperature was attached to the transmitter, outside the fish's body on the harpoon shaft. For other experiments, in which we measured the stomach temperature, an instrument package was pushed down the bluefin's throat and into its stomach; the thermistor that measured the water temperature was at the end of a wire we led out through the tuna's gill slit. The transmitters broadcast the temperature of the fish and the temperature of the water during alternate one-minute periods.

Our tracking vessel was equipped with a directional hydrophone that enabled us to follow the tuna. We rotated the hydrophone until the telemetered signal was at a maximum and then steered the vessel on that heading. At hourly intervals we lowered a bathythermograph to measure the temperature of the water at various depths. By comparing these readings with the telemetered temperature of the water surrounding the fish we could estimate the bluefin's swimming depth, and we could approximate the fish's course by plotting our own course as we followed the signals.

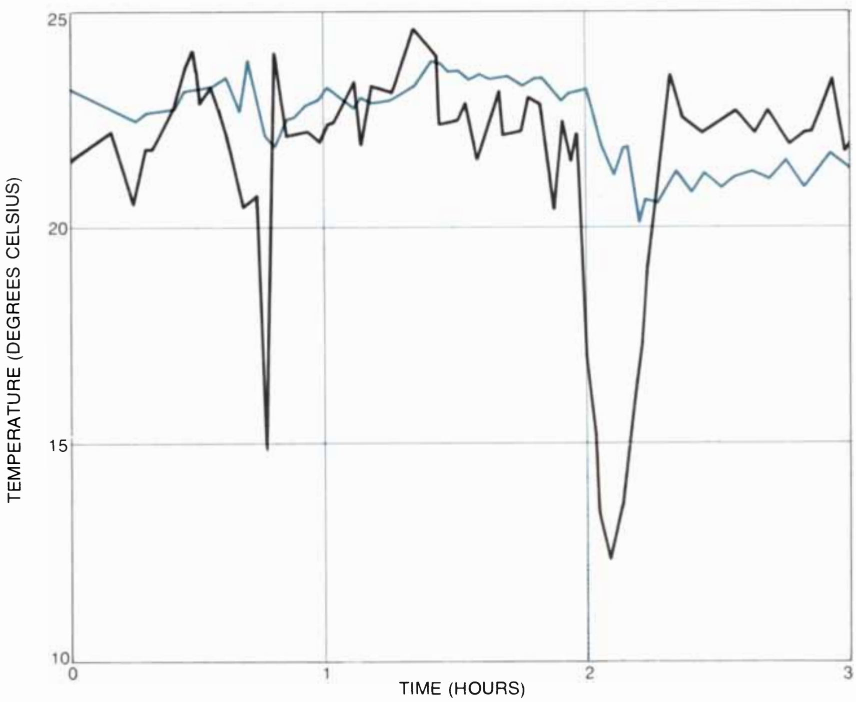
We placed transmitters on or in 14 bluefin tuna. The longest we tracked a fish was 54 hours; by then the bluefin had reached a position 130 miles offshore. We soon found that, as commercial fishermen had told us, tuna avoid changes in water temperature if they can. Most of our specimens remained near the surface or at least on the warm side of the steep thermocline that separated the upper waters from the cold depths. Some of the bluefin would dive through the thermocline, but they spent only a few minutes in the colder water before returning to the warm side.

Near the end of our efforts we were lucky enough to get a most satisfactory result. The specimen was a 600-pound bluefin with a transmitter in its stomach. The fish had been handled quite roughly while the instrument was being installed. Perhaps for this reason as soon as it was released from the pound at about 9:00 A.M. it swam down through the thermocline into water 14 degrees C. colder than the surface water in the pound. When the fish was released, its stomach temperature registered 21 degrees C. During its four-hour stay in the five-degree water the temperature of its stomach gradually fell to about 19 degrees. Early in the afternoon the fish returned to the warm side of the thermocline and remained in water that registered between 13 and 14 degrees for the rest of the day. In spite of a change in water temperature of nearly 10 degrees C. the temperature of the fish's stomach remained around 18 degrees. The fact that the deep-body temperature of the fish remained nearly constant over extended periods in both cold water and warm indicates that the bluefin was indeed thermoregulating. Just how the fish do this we do not know. Presumably it is by somehow varying the efficiency of the heat-exchanging *retia.*

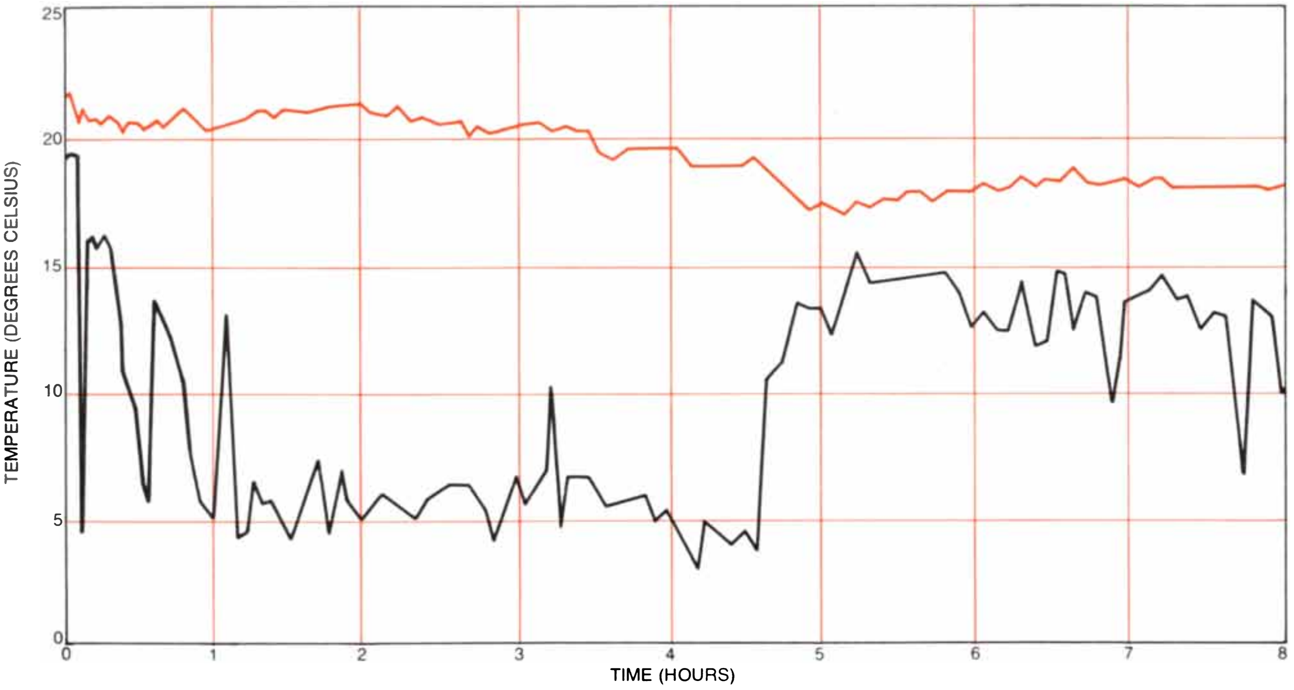
We were not fortunate enough in our muscle-temperature experiments to have a bluefin stay on the cold side of the thermocline for any substantial length of time. One of the muscle-tagged fish, however, did swim for hours in water that was gradually decreasing in temperature. The water temperature dropped four degrees C. in one 90-minute period. During this interval the muscle temperature of the bluefin slowly rose from a



TELEMETRY RECORD compares the temperature in the muscle of a bluefin (*color*) with that of the water (*black*) through which it swam for three hours. The muscle temperature was held constant as the water temperature declined gradually from about 14 to 10 degrees.



RECORD OF TEMPERATURES from the muscle of a dusky shark (*color*) and the water it swam in (*black*) shows a different relation. The dusky shark is a cold-blooded fish and its muscle temperature stayed close to that of the water, dipping during even short dives.



RECORD FROM TUNA'S STOMACH shows the effect of the visceral, as opposed to the muscular, temperature-control system. The temperature was telemetered from thermistors in the tuna's stomach (*color*) and outside the fish in the water (*black*). When the tuna was released, it swam down into cold water and stayed there for four hours, but its stomach temperature decreased only a little.

little more than 23 degrees C. to above 24 degrees.

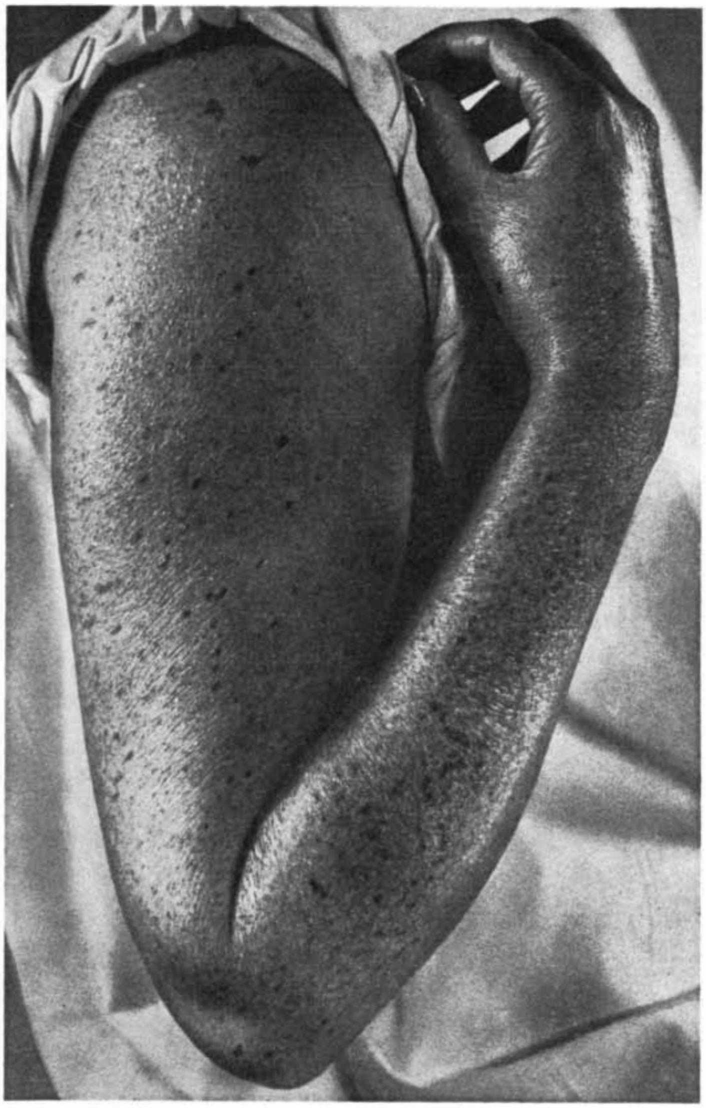
Our experiments showed that the bluefin does not maintain its body temperature at a constant level as, for example, human beings do. The temperatures of fish from the same school can fluctuate over a range of five degrees or so. The greatest fluctuation we recorded was provided by a second tuna that carried a stomach temperature transmitter. At the time of release the temperature of the fish's stomach was 19 degrees C. During the first day of tracking its stomach temperature slowly began to rise. By the second day, although the water temperature had remained constant, the fish's stomach was registering 26 degrees C. Perhaps the fish was trying to digest the transmitter!

In other readings we found that muscle temperature tended to cycle over a narrow range that is unrelated to the temperature of the water. All these variations tell us that whereas the bluefin's various organs are being kept much warmer than the surrounding water, precise temperature control is either not necessary or not possible. In effect, the fish has a sloppy thermostat. Nonetheless, these fish can be said to thermoregulate in the same sense that birds and mammals do.

How did the tunas and mackerel sharks come to develop these remarkably efficient heat exchangers? Although the *rete* is certainly a complex vascular array, its mode of evolution was probably quite simple. For one thing, arteries and veins generally follow the same path as they travel between organs and within tissues. Every such pairing of an artery and a vein is a rudimentary heat exchanger; if at one end of the system there is either heating or cooling, then there will be some heat transfer between the paired blood vessels. This rudimentary exchange is probably what accounts for our observation that fishes that ought not to be able to warm themselves at all by metabolic means nevertheless may have muscle temperatures from one to two degrees C. higher than the water temperature.

Suppose that such a slight rise in tissue temperature offers a genuine selective advantage to an evolving species. In such an event there already exists (as part of the normal embryonic development of the circulatory system) a basis for the development of more advanced heat exchangers. The circulatory system of the developing embryo first takes the form not of discrete blood vessels but of beds of interconnecting spaces and channels. Later, as certain channels become important, these routes become larger and other channels atrophy and disappear. Moreover, there is a tendency toward the development of multiple channels. Several such channels may form within a single bed, with the result that the vascular system of the adult will have some duplicate components. Therefore in order to form the mass of parallel blood vessels that is an efficient heat exchanger, only a modification of the normal embryonic pattern is required, not a radical genetic change.

A fast-swimming predatory fish finds an abundance of food awaiting it in the form of fast-swimming squid, herring and mackerel that slower predators cannot capture. In terms both of streamlining and of muscle power the tunas are probably the swiftest predators of the open ocean. Yellowfin tuna and wahoo tuna are noted for their speed: they have been observed to reach speeds in excess of 40 miles per hour during 10- to 20-second sprints. Mackerel sharks (at least the mako and porbeagle) probably also enjoy the benefits of fast swimming. Such a shark, with its bulky, muscular body, its streamlining and its narrow, crescent-shaped tail, looks remarkably like a tuna. That the two unrelated groups of fishes independently evolved the same means of raising their body temperature must surely be connected to the adaptive advantage of increased swimming speed; the extra power available from warm muscle must have been decisive in achieving that speed. It is a classic example of parallel adaptations that evidently gave both groups access to an underexploited source of food.



Kodak

Armed with experience

In the cause of objectivity

Two of the three kinds of lighting are used in this example of medical photography: contour lighting to show the well-developed arm and forearm, texture lighting to depict the characteristics of ichthyosis vulgaris. Flat lighting is the third kind, and just as important in medical photography. Lighting is often the chief component of photographic style, but a glamorous photographic style is out of place in clinical photography.

Many shops that carry extensive lines of photographic equipment and supplies stock the new KODAK *Data Book "Clinical Photography" (cover price, $2.95). In a mere 118 thoroughly illustrated pages it summarizes decades of personal experience and consultation in the field by a recently retired Kodak man who once served his stint as president of the Biological Photographic Association. Assuming some understanding of cameras and sensitized materials, he writes to would-be medical photographers and medical people of all ages who strive to make the practice of medicine ever more objective.*

Sharing chemical thought

Ask ten reasonably alert citizens at random for their conceptions of what a research chemist does at work. Then ask ten research chemists how they spend their working day. Compare. Note that the real eyestrain comes not from staring at test tubes and instrument dials but at publications, reports, and patents from other chemists.

To reinvent the wheel is humiliating, wasteful, and (if the patent on the particular wheel hasn't expired) financially perilous. Hence the rise of chemical documentation, an underappreciated discipline not lacking in intellectual challenge of its own.

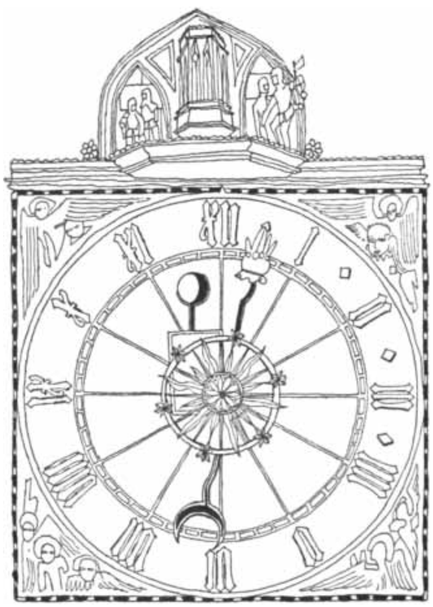
As a major investor in chemical research, Kodak does appreciate chemical documentation and has also invested quite a lot in the development of *that* discipline. Our returns on the latter investment safeguard our return on the former. Our contributions to chemical documentation stress "browsability" among structures. Superelegant, precise computer output must serve mere humans whose ideas of what they are searching for change even as they search.

We see at least two ways this could serve chemists other than our own or our industry's:

1. We are in the microfilm and microfilm equipment business. Got it going, in fact. Back then in 1928 it was hardware and film. Nobody thought of them as mere tools of something called "information technology." Today information technology has few worthier tasks than to keep chemistry from sinking of its own weight. Our thinking in this direction has had to go deep beneath the generalities. It may be worth sharing.

2. We also conduct an entirely different kind of business in custom production of compounds not generally available beyond laboratory quantities. In serving prospective customers of *that* business, our chemical information resources are no less important than equipment and the skills to operate it.

Inquiries in either area can be addressed to Kodak, Dept. 55W, Rochester, N.Y. 14650—the more specific the more welcome. Requests merely to "send literature" will only mystify us.



SCIENCE AND THE CITIZEN

Leveling Off

The reproductive behavior of Americans appears to have changed significantly, with important implications for the rate of growth and the future structure of the U.S. population and economy. The fertility rate has been falling for some time; so have the crude birthrate and the actual number of births, in spite of an increased number of women of childbearing age. Perhaps most significant, recent statistics show that the total fertility rate—the projected size of completed families—has dropped below 2.1 children per family. That is the level at which the population can eventually attain a steady state, with births just balancing deaths. The present age structure of the population and the likelihood of changes in fertility patterns mean that the U.S. will not soon enter a period of zero population growth, but the Bureau of the Census is apparently convinced that population growth will be slower than had been expected: its latest projections, issued in December, of the population in the year 2000 have been revised sharply downward from the 1971 estimates.

The U.S. birthrate was 23.7 per 1,000 of population in 1960, 17.3 in 1971 and only 15.8 in the 12 months ending last October. The general fertility rate (births per 1,000 women of childbearing age, or 15 through 44) was 82.9 in the first 10 months of 1971 and only 73.6 in the same period last year. As a result it appears that the total increase in the U.S. population last year was only about .7 percent as compared with an average annual increase of 1.1 percent from 1963 to 1970. The total fertility rate is more clearly related to family size. It is derived by summing the age-specific rates for the 30 childbearing years, and thus tells how many children a hypothetical woman would bear if she experienced the current fertility of women at successive ages. (The zero-growth rate is 2.1 rather than 2 per family because some girls die before reaching the age of childbearing.) The total fertility rate was more than 2.2 even during the Depression. It rose to 3.8 in 1957, declined to 2.46 in 1968 and to 2.39 in the first nine months of 1971. In the same period last year it dropped to 2.08.

The reason appears to be primarily that more young women are not marrying, are marrying later or are marrying and nevertheless delaying having children; this could indicate a long-term change in reproductive behavior or it could merely cause a temporary decrease in population growth. Even if the average family size were stabilized at 2.1, the population would grow from its present 209 million to about 320 million in some 70 years before leveling off, because a disproportionately large number of women will be reaching childbearing age in the years ahead.

The fact remains that all current trends point to a period of slower population growth. Notable among these trends are the increase in the proportion of women who work and the expressed intention of women at all economic and educational levels to have smaller families, an objective that will be easier to attain because of the increased efficacy of contraceptive techniques and the increased availability of legal abortion. As a result the Bureau of the Census has modified its population estimates for the year 2000. The bureau makes four projections on the basis of various total fertility rates and it has eliminated its previous "high" projection, based on a rate of 3.1, of 322.3 million. The current high projection, based on a rate of 2.8, is a U.S. population of 300.4 million in 2000. The next two projections, based respectively on rates of 2.5 and 2.1, are 286 million and 264.4 million. A new "low" projection has been introduced based on a rate of only 1.8, which would generate a population of 250.7 million in 2000. As recently as 1967 the highest projection was 361 million and the lowest was 280.7 million.

Secrets

In the years since the declassification of the U.S. controlled-thermonuclear-research program in the late 1950's the quest for fusion power has been the very model of international cooperation in scientific investigation. Representatives of the leading U.S. and Russian fusion-research laboratories, for example, have regularly exchanged visits, results of experiments and even helpful advice. This policy of openness apparently does not apply to one of the newest branches of the controlled-thermonuclear-research effort: the attempt to induce power-releasing fusion reactions between hydrogen nuclei by bombarding a small pellet of frozen hydrogen with a pulse of light from a high-powered laser. Presumably because such work is thought to have some military potential (perhaps in the design of a "clean" all-fusion bomb or a "death ray" ABM system) laser-induced fusion has been subjected to classification procedures reminiscent of the early postwar atomic-bomb period.

Commenting recently on "the new secret subject," S. A. Goudsmit, an editor of *Physical Review Letters,* observed: "We don't know why it [laser-induced fusion] is secret. We have been told to mind our own business and not mingle in matters of Government policy or politics. Our own business is publishing papers on advances in physics, and that is where we got into difficulties with secrecy."

Goudsmit, who made his remarks in a New Year editorial in *Physical Review Letters* titled "Secrecy Again," continued: "All we know about fusion is what we read in the newspapers and journals. It is heralded as the energy source which will solve all the world's problems. It is inexhaustible and pollution-free. The remotest corners of the world will get enough energy so that everybody will be rich and can relax. Keeping work towards this goal secret sounds like hiding progress in cancer research. Perhaps it is done to prevent raising false hopes, or to prevent the premature sale of stock in oil companies. We don't know, and we don't understand."

In an effort to demonstrate the fundamental incompatibility of scientific progress and a policy of secrecy in reporting scientific results, Goudsmit describes a recent incident involving his journal. Having received a communication on the subject of laser-induced fusion, the editors of *Physical Review Letters* sent it to a referee, who prepared a long critical report. The referee, however, could give the journal only two short paragraphs, "since the rest has to be declassified, which will take a couple of months." When the author of the original letter was informed of the delay, he responded that he was "not surprised" and that the referee's critique most likely concerned items that the author had to omit from his letter in the first place because of secrecy.

"Can physics advance in this way?" Goudsmit asks. "Isn't it silly!" He concludes his editorial with the hope "that the New Year will bring some enlightenment."

Creation Compromised

Divine creation will apparently not be taught in California as an alternative to evolution after all, at least not in natural-science classes. After some years of controversy the State Board of Education has decided that creationism need not be given equal treatment with biological evolution in elementary-school textbooks. Instead, the board decided, social-science textbooks will have to include discussions "concerning the representative philosophies of the origin of man."

The decisions represented a compromise. Fundamentalists, working primarily through an organization called the Creation Research Society, have been arguing that evolution should not be accepted as "dogma" and that special divine creation should be put forth as at least a respectable alternative. Organizations of scientists and science teachers, local school boards and representatives of the major religions have contested that view. In December the State Board voted down a proposal that would have required equal treatment for creation in natural-science books, but it called for changes in such books to edit out "scientific dogmatism" concerning evolution. Last month the board again rejected (by a six-to-three vote) a proposal to insert creationism in natural-science books but mollified the creationists by requiring treatment of the subject in social-science texts. The editing out of "scientific dogmatism" in natural-science texts continues, however. Moreover, the board has yet to face the question of just how, and how extensively, religious theories of creation are to be written into social-science books covering subjects from history and geography to economics, psychology and anthropology.

Missing Pulses

It has now been three and a half years since Joseph Weber of the University of Maryland first reported that he had detected pulses of gravitational radiation coming from space [see "The Detection of Gravitational Waves," by Joseph Weber; SCIENTIFIC AMERICAN, May, 1971]. His results generated wide interest, since they promised a new source of information about the universe, and since the pulses he was detecting implied the occurrence of cataclysmic events in the center of the galaxy. Several groups of workers around the world undertook similar efforts. One of the more ambitious experiments, conducted by J. A. Tyson of Bell Laboratories, was discussed at the Sixth Texas Symposium on Relativistic Astrophysics, held in New York late in December. So far the results of Tyson's experiment are negative.

Weber's apparatus was basically a solid aluminum cylinder five feet long and three feet in diameter. Attached to the surface of the cylinder were piezoelectric crystals that could detect changes in the cylinder's dimensions equivalent to as little as a millionth the diameter of an atom. Sudden changes in dimension were assumed to be due to the passage of pulses of gravitational radiation. One cylinder was located at College Park, Md.; another was set up 600 miles away at the Argonne National Laboratory outside Chicago. The two instruments were connected by a high-quality telephone circuit so that pulses detected simultaneously by both of them could be counted. As many as three coincidental pulses per day were reported. The data seemed to be correlated with times at which the rotation of the earth swept through the center of the galaxy a plane perpendicular to the cylinders' long axes.

Tyson began observing last year with an antenna that would be some 200 times more sensitive to pulses of gravitational radiation than Weber's antennas were in 1970. The apparatus is a cylinder 12 feet long and 27 inches in diameter located in a concrete bunker in the middle of a swamp near Holmdel, N.J. In the three months that this larger antenna has been "on the air" it has not detected a single pulse.

According to Tyson, the explanation may be that the source of gravitational radiation is variable and is not currently radiating. It may also be that many or all of the coincidences observed with the Weber antennas do not originate with pulses of gravitational radiation. The coincidences seem real, but except for their apparent correlation with the passage overhead of the galactic center they could be explained by phenomena closer to home. For example, a shock wave in the "wind" of particles traveling outward from the sun could create a disturbance in the earth's magnetosphere that could be picked up simultaneously by two widely separated detectors.

If the pulses detected by Weber's apparatus are gravitational, each pulse would require for its generation an amount of energy equivalent to the amount that would result from the total conversion into energy of a body with the mass of the sun. It seems likely that such a dramatic event would also give rise to other forms of radiation. Therefore radio waves, infrared radiation and X rays from the galactic center have been monitored to see if there are bursts of these other forms of energy that can be associated with the coincidences attributed to gravitational radiation. So far no such bursts have been detected.

At least 14 workers (or groups of workers) in addition to Tyson are currently trying to detect gravitational pulses or are expecting to have antennas on the air within the next year or two. It seems likely that the question of whether or not pulses of gravitational radiation have been detected will be resolved in that time.

(Crankcase) Oil on the Waters

What is the major source of the oil with which man pollutes the ocean? It is not tanker accidents, clandestine bunker-flushing, sea-floor drilling disasters or similar well-publicized occurrences. More than two-thirds of the oil comes from the disposal of billions of quarts from the crankcases of automobile engines and the oil sumps of other machines around the world each year. That is the conclusion of a report to the Connecticut state legislature based on data from the Conservation Division of the U.S. Geological Survey and the Environmental Protection Agency.

The report, prepared by Mallory Factor, finds that some legislative curbs are beginning to be placed on pollution by waste oil from industrial sources. The scant legislation that exists with respect to disposal of used motor oil, however, is largely unenforced. One measure of the problem is that whereas a decade ago commercial collectors delivered some 300 million gallons of waste motor oil to U.S. refineries every year, today only 100 to 150 million gallons of the 500 million gallons of oil annually drained from the nation's crankcases are re-refined. Most of the rest, the report concludes, is dumped into sewers and waterways and eventually ends up "polluting the waters of the world."

The Paranoid Computer

DOCTOR: "How long have you been in the hospital?"

PATIENT: "About a week."

DOCTOR: "Are you depressed?"

PATIENT: "No."

DOCTOR: "Were you taking drugs?"

PATIENT: "No."

DOCTOR: "What is your problem?"

PATIENT: "People make me nervous."

DOCTOR: "How do they do that?"

PATIENT: "They stare at me."

DOCTOR: "Why?"

PATIENT: "Maybe they don't like the way I look or something."

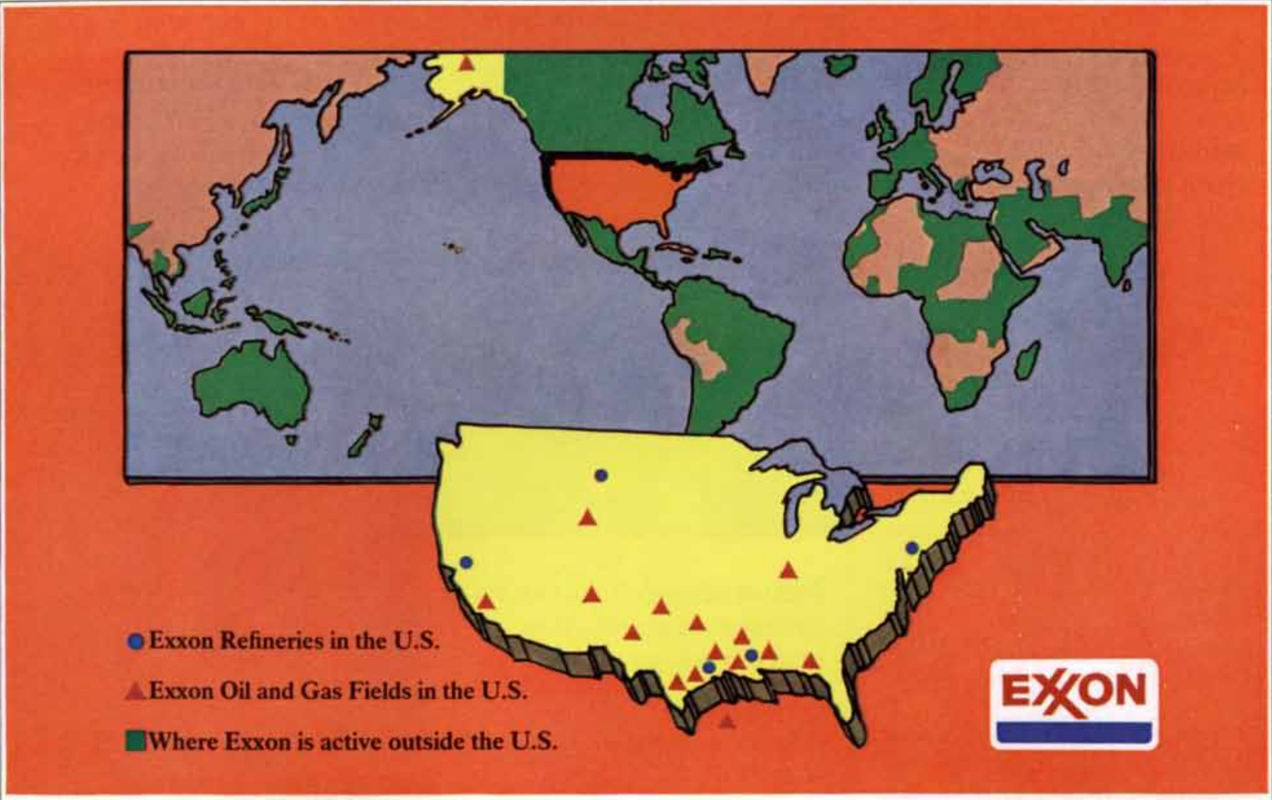
DOCTOR: "Do you look unusual in some way?"

PATIENT: "Why are you interested in my looks?"

This dialogue is the first part of an actual diagnostic interview between a psychiatrist and a patient who has systematized delusions of persecution. Unknown to the psychiatrist, however, the patient is a computer programmed to simulate paranoid processes in human beings. The hypothetical patient is a young male postal worker who lives alone and is afraid that "the underworld is out to get him" because he has assaulted a bookmaker during a quarrel. Out of 25 psychiatrists who interviewed the computer model, 23 judged the patient to be paranoid. Two considered it to be brain-damaged because of its linguistic limitations.

The simulation of paranoia is being conducted by Kenneth M. Colby, Franklin D. Hilf, Sylvia Weber and Helena C. Kraemer of the Artificial Intelligence Project at Stanford University. They believe that a successful theoretical model of paranoia will embody as part of its inner structure an explanatory account for certain empirical regularities found in paranoid patients. Several studies have shown that there is very high agreement, from 85 to 95 percent, among psychiatrists in rating the presence and degree of paranoia in a patient.

Weak and strong versions of the model were developed to see if the degree of paranoia exhibited by the model corresponded with the judgments made by



A corporation should be a "corporate citizen" as well.

The people of Exxon Corporation have made sure that the company is committed to goals that go beyond the primary responsibilities of conducting their business efficiently. There is also a commitment to society, and that commitment is an integral part of the business of Exxon.

The Exxon Education Foundation, for example, has a seventeen-year history of financial support to higher education—in fact, it has contributed over 34 million dollars to date.

And for nearly as long, Exxon has supported fine arts through competitions around the world, educational experiments like Harlem Prep, and excellence in television programming like *An Age of Kings* and *Play of the Week* and, more recently, *Vibrations*.

Beyond this are the individual activities of our many employees working, often with Exxon support, in community projects as varied and numerous as the skills they offer.

We'd like you to know Exxon.

We hope these pages have told you something about Exxon Corporation—who we are, what we do and where we work.

We are U.S.-based with many affiliates abroad. There are, in fact, 143,000 Exxon employees around the world. Our primary task is finding, producing, transporting, refining and selling oil and natural gas to millions of customers.

In the United States, there are 40,000 of us; most of us work for Exxon Company, U.S.A., in itself one of America's leading oil and gas enterprises. In recent months Exxon Company, U.S.A. has carried out the change to Exxon that you've seen at your service station.

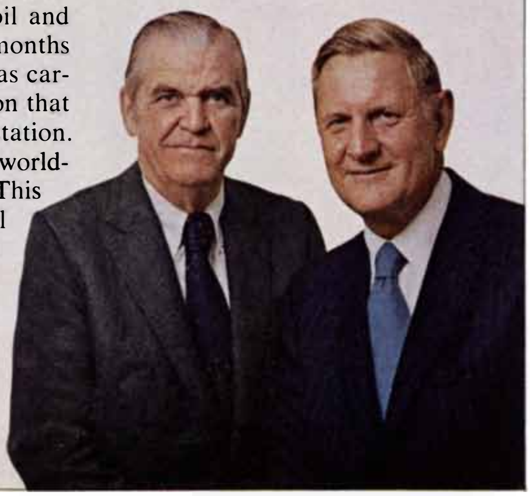
An important part of our worldwide business is chemicals. This is guided by Exxon Chemical Company. U.S. chemical operations are conducted by Exxon Chemical Company U.S.A.

We are also companies that do research and engineering, international exploration, development of non-petroleum energy sources, and land development. This should give you an idea of our scope.

Although the name of Exxon Corporation is new, we realize that our reputation will continue to depend upon the quality of our products, the service we offer and the way we conduct ourselves.

J. K. Jamieson
Chairman,
Exxon Corporation

C. C. Garvin, Jr.
President,
Exxon Corporation



the psychiatrists. In the interview the psychiatrist communicated with the patient by means of a teletypewriter. He interviewed two patients one after the other. In half of the runs the first interview was with a human patient and in half the first interview was with the computer model. Four paranoid patients from a hospital participated in the experiment. In addition the investigators sent transcripts of the interviews along with a third interview with a nonparanoid patient (also conducted with a teletypewriter) to outside psychiatrists for evaluation.

Both the psychiatrists conducting the interviews and the outside psychiatrists judged the strong version of the model to be more paranoid than the human paranoid patients and the weak version to be less paranoid. Virtually all the ratings for the nonparanoid patient were zero. Colby, Hilf, Weber and Kraemer conclude that the results support the claim that the model is a good simulation of a paranoid patient. To test the explanatory power of the simulation the next steps, they say, would be for the model to predict properties of paranoia not yet discovered in patients. Another strong validation of the model would result from successful therapeutic attempts to change the model that could be duplicated in the treatment of paranoid patients.

Fact

Modern man is addicted to superlatives: the highest mountain, the deepest deep, the largest animal that ever lived and so on. Is Mount Everest really the highest mountain? This widely accepted fact is called into question by Robert L. Birch, a reader of SCIENTIFIC AMERICAN. According to Birch, the highest mountain is probably Chimborazo in the Ecuadorian Andes.

It all depends on what is meant by "high." Mount Everest rises 29,028 feet above sea level, Chimborazo 20,561 feet. It is well known, however, that since the earth is an oblate spheroid, the sea level at the Equator is some 14 miles farther from the center of the earth than the sea level at the North Pole. Indeed, distance from the center of the earth would seem to be a more reasonable standard for the height of a mountain than sea level is, since it is a measure of how far the mountain sticks out into space. On this basis Chimborazo, which stands within two degrees of the Equator, is some two miles higher than Everest, which is nearly 28 degrees from the Equator.

THE CHINESE LANGUAGE

This melodious tongue is spoken by more people than any other. Although the Chinese system of writing is complex, the basic structure of the language is remarkably simple

by William S-Y. Wang

To people who are familiar only with the common European languages the Chinese language is strikingly different. Yet today Chinese is spoken by more people than any other language, and Chinese literature is the world's oldest, spanning a period of 35 centuries. When we examine the structure of the Chinese language, we find that it is not conspicuously complex; indeed, in many ways it is simpler than the Western languages. But since Chinese does differ from the European languages in fundamental respects, some knowledge of its structure and historical development is indispensable to a general understanding of the nature of human language.

To the Western eye the writing system of the Chinese is altogether novel: instead of neat rows of simple alphabetic letters there are thousands of unique characters, many of which seem incredibly intricate. To the ear the language sounds rather melodious, perhaps a little like singing. When one peers below the surface, there are more surprises. The language has virtually no conjugation for its verbs and no declension for its nouns. The inevitable paradigms that Western schoolchildren have come to dread in their grammar books are totally absent in a grammar of Chinese. For the various forms of the verb "to buy," such as "buy," "buys," "bought" and "buying," Chinese has the single form *mǎi*. (The mark over the *a* signifies that the syllable is spoken in a tone that falls and then rises. There are three other vowel marks for tones, as in *á* indicating a rising tone, *à* indicating a falling tone and *ā* indicating a level high tone.) For "book," "books," "to the books" and "of the books" the Chinese is *shū*. Most of the time it is quite clear from the context what tense or mood is intended for a verb and what number or case is intended for a noun. Hence the Chinese language does not bother much with this particular type of redundancy in its grammar. Perhaps it is this structural simplicity of the language that moved the anthropologist and linguist Edward Sapir to characterize it as "soberly logical."

SYMBIOTIC RELATION between painting and writing in Chinese art is elegantly exemplified by a detail from "Flowering Plants and Trees" on the opposite page. The characters were written in the traditional format: in columns and from right to left. The portion of the poem that is framed by the pine branches is translated: "Jade strands hang limp in the wind. Crimson berries sparkle bright against the snow." The painting was made by Ch'en Shun in the 16th century during the Ming dynasty. It is part of the Avery Brundage Collection at the Center of Asian Art and Culture in San Francisco.

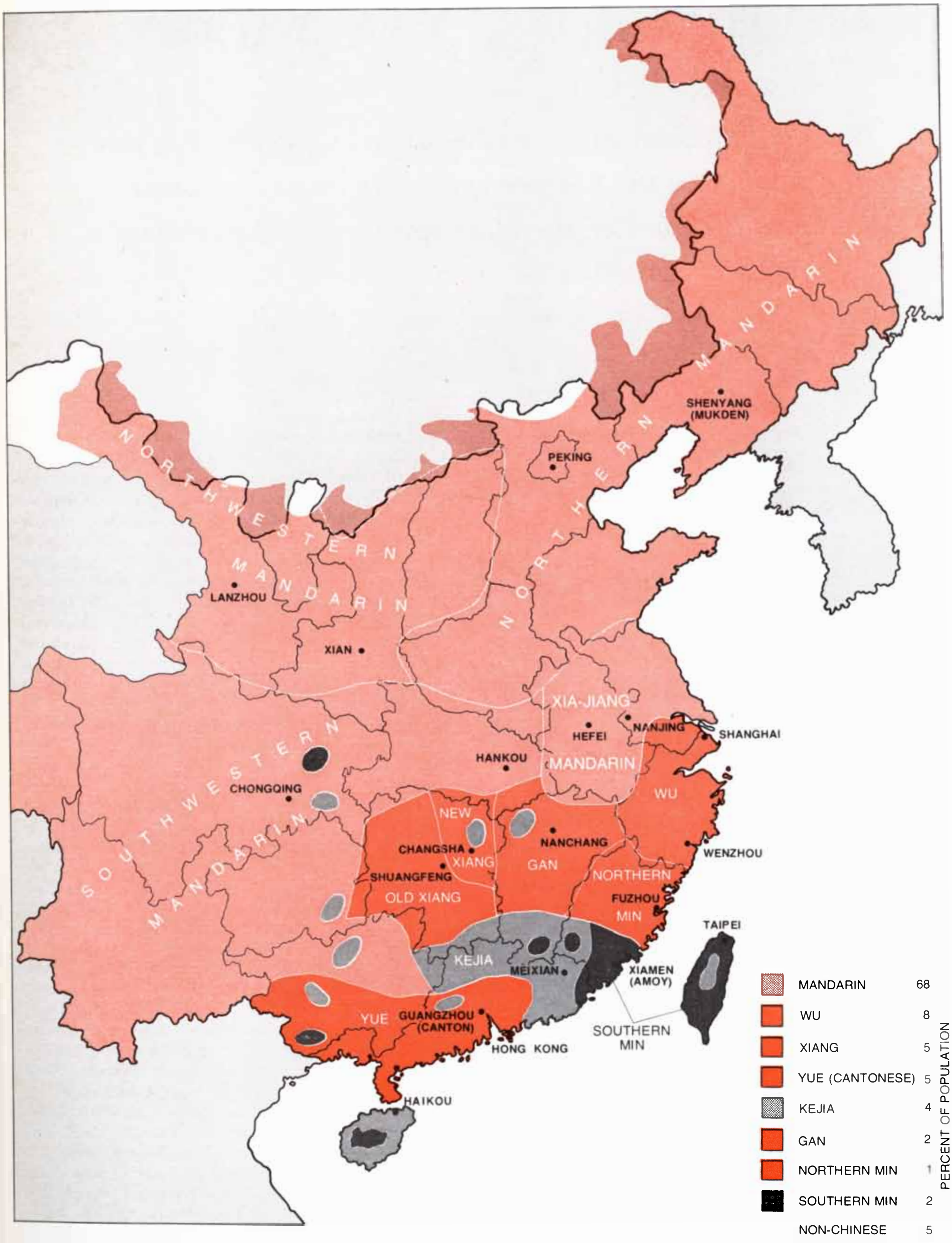
The Antiquity of Chinese

Chinese is often termed a very old language. In a sense such a statement is misleading. All human languages go back to the dim uncertainty of prehistory, and at present we have no way of knowing whether or not they can all be traced back to the same root. Four thousand years ago the ancestors of the Chinese peoples spoke an early form of the Chinese language in much the same way that the ancestors of the English-speaking peoples were using an early form of the English language. Since almost nothing is known about the emergence of language in the human species, we are not in a position to say which of the world's languages evolved earlier and which later. It is rather that in the course of history some languages have been renamed more often than others (as a result of events such as migration or conquest) and the newness of the names gives the illusion that the thing being named is new.

There is one sense, however, in which Chinese is a very old language. Sumerian is the only language we know of that has extant written materials that antedate Chinese ones. Sumerian cuneiform writing dates back some 5,000 years; the earliest Chinese writing in existence today dates back 3,500 years. But Sumerian and its derivative orthographies died out long before the beginning of the Christian Era. Chinese orthography has continued to this day, although there have been major stylistic changes.

The earliest Chinese writings are incisions on bone and tortoise shell. Most of the inscriptions are oracular, dealing with political or religious events or with the weather or warfare. Discovered toward the end of the 19th century in Chinese drugstores, where they were being sold as "dragon bones" for their medicinal value, the story of these inscriptions is a colorful chapter in the history of Chinese archaeology and philology. More than 100,000 inscribed pieces have now been found. Even though the total number of written characters on the pieces is more than a million, the number of different characters is small. The texts of the oracular inscriptions dealt with a very limited range of topics, and the same characters are repeated over and over again. Of the 2,000 to 3,000 characters on the shells and bones, about half can be read today.

Through the centuries Chinese characters have been preserved in many different mediums: metal vessels, stone drums, jade jewelry, coins, metal mirrors, bricks and tiles. The central line of development, however, has been the use of the brush on silk, bamboo, wood and ultimately on paper. A brush can produce variations in thickness whereas a stylus cannot. Such variations give the



MAP OF CHINA shows the distribution of the major dialects of the Chinese language. More than two-thirds of the Chinese population speak one of the Mandarin dialects, of which the speech of Peking is the standard. It is the dialects along the southern coast, however, that have been carried to many parts of the world by Chinese emigrants. There are also several non-Chinese linguistic stocks within China. The regions to the north and west are dominated by non-Chinese languages such as Mongolian and Tibetan.

writer a much greater artistic freedom in rendering his characters.

Some of the earliest written Chinese characters were pictographic. The character for "rain" was several columns of broken lines, and the one for "horse" looked like a horse, complete with mane and four legs [*see illustration on page 59*]. Pictographs, however, are only a minority in the total vocabulary of Chinese. Most of the words in the language cannot be suggested by a simple picture.

Calligraphy, the elegant rendering of characters, is a highly cultivated art form that has long been prized in Chinese culture, much as painting is valued in the Western world. For the Chinese there is a close relation between painting and calligraphy. Typically a silk scroll is covered with a picture and a few lines of characters, the two carefully balanced against each other. Because of their artistic values and their long history, Chinese characters have a much greater range of variability in their size and shape than the characters of any other writing system.

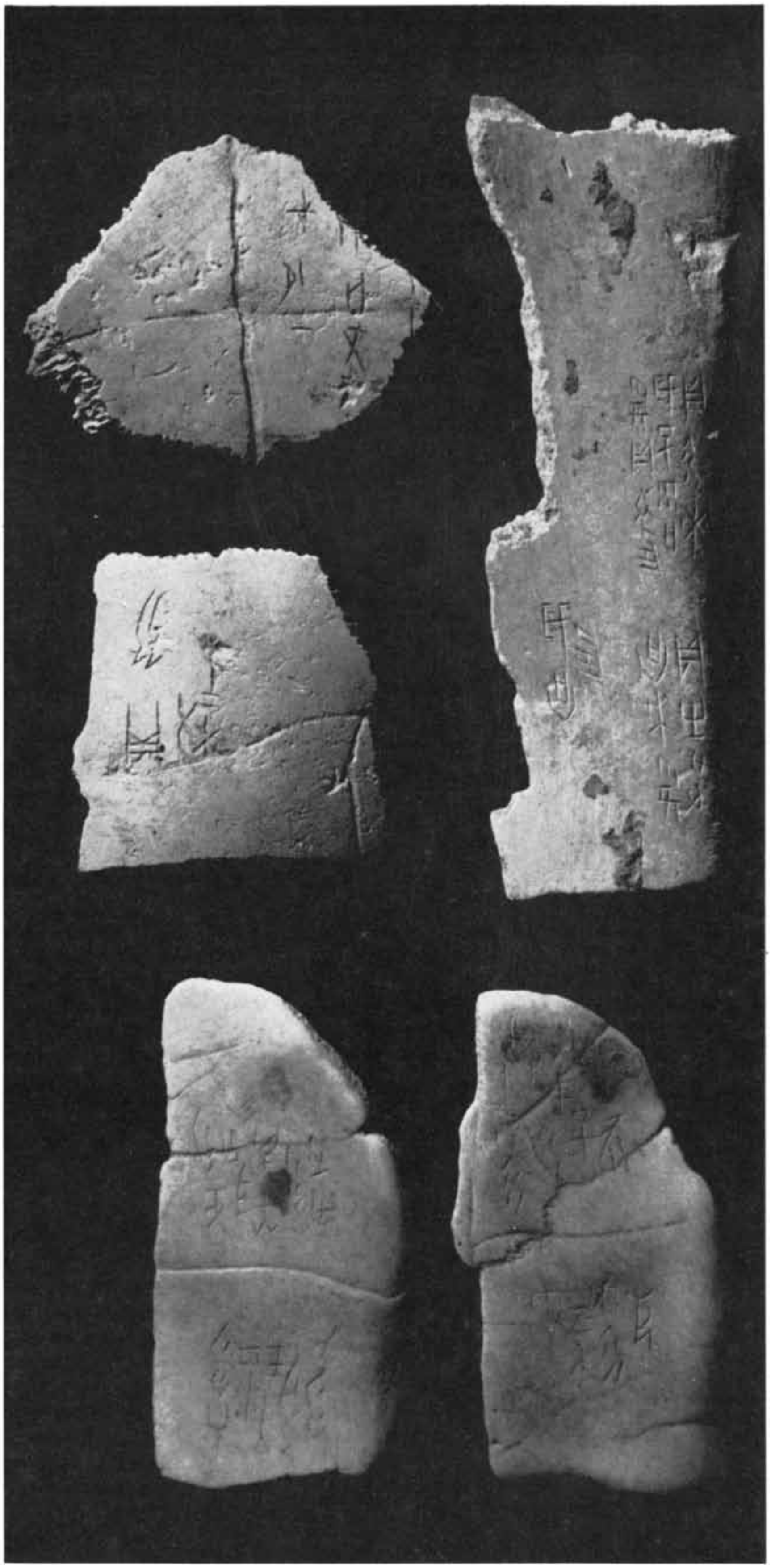
The Writing System

The Chinese writing system underwent major changes in 1956, when the government of the People's Republic of China decided to simplify the characters and also to adopt a system of spelling Chinese words in Latin letters. Both measures are intended to make the reading and writing of Chinese easier to learn, a crucial step in promoting linguistic unity and raising the standard of literacy in China.

In order to understand the nature of Chinese characters and their simplification, we must first examine their internal structure. Each character is made up of two types of smaller unit called the stroke and the radical. Roughly speaking, a stroke is a line, either straight or curved, that is completed every time the pen leaves the paper. For example, the character for "sun," which is pronounced *rì*, looks like:

日

It is built up of four strokes [*see top illustration on page 55*]. Both the order and the geometric position of the strokes are important. There are approximately 20 distinct strokes in the language, so that strokes are the closest counterparts to the 26 letters of the Latin alphabet. There is no counterpart of the radicals in the orthography of other languages. The traditional set consists of 214 radicals, and these radicals are found in almost all



ORACLE-BONE INSCRIPTIONS, dating from 1300 B.C., are among the earliest-known examples of Chinese writing. They were made on ox bones and tortoise shells and were used for divination. The photograph, provided through the courtesy of Bernhard Karlgren and Jan Wirgin, shows specimens in the Museum of Far Eastern Antiquities in Stockholm.

Chinese dictionaries. A dictionary published in 1971, however, has merged some radicals, reducing the number to 189. Most radicals are also characters; for example, the "sun" radical and the character for "sun" are identical. There are thousands of characters in regular use, and the majority of them are not radicals. Each character contains only one radical, with or without a remainder.

The character for "star," pronounced *xīng*, is written:

星

There is a literary character referring to the morning sun, pronounced *lóng*, that is written:

昽

Now it can be seen that both characters contain the "sun" radical:

日

"Star" has the radical on top and a remainder, which is pronounced *shēng*, that looks like this:

生

"Morning sun" has the "sun" radical to the left; its remainder is pronounced *lóng* and is written:

龙

Almost all the recent Chinese dictionaries and reference books arrange their characters by radicals. Take for example the character for "star." To look it up one first goes to the section marked by the "sun" radical:

日

In this section all the characters that contain the "sun" radical are ordered by the number of strokes in the remainder, which is:

生

This remainder is only moderately complex, with five strokes. We would expect to find it in about the middle of the section.

A key process in the construction of characters can be illustrated with "star" and "morning sun":

星 昽

Their meanings are clearly related to the meaning of the "sun" radical. In these characters the radical is called the signific and the remainders are the phonetics. The phonetics indicate how the characters are to be pronounced. In "morning sun" the phonetic is *lóng:*

龙

It gives its pronunciation to the character. On the other hand, there are characters that are not pronounced like their phonetic, often for reasons of historical sound change. The phonetic in the character "star" is:

生

The phonetic is pronounced *shēng*, but the pronunciation for "star" is *xīng*.

As another example of a phonetic we can take the character for "horse." It is pronounced *mă* and is written:

马

When the phonetic for "horse" is combined with the signific for "woman," we have *mā*, which means "mother":

妈

When the "horse" phonetic is combined with the signific for "jade," we have *mă*, which means "agate":

玛

Combined with the signific for "insect," the meaning becomes "ant," and again it is pronounced *mă:*

蚂

When there are two "mouth" significs hovering over the "horse" phonetic, the meaning becomes "to scold," and it is pronounced *mà:*

骂

There are a large number of characters in Chinese that are constructed on the phonetic-signific plan. Thus underlying many Chinese characters there is a phonetic principle. The average Chinese can often pronounce correctly a character he has never seen before simply by making a shrewd guess at its phonetic. For example, examine the following:

The signific portion means "gold." The phonetic portion is pronounced *lóng* and is written:

龙

There would be no problem in agreeing that the character should be the name of a metal or metallic compound and should be pronounced like *lóng*, even though such a character does not exist in the language.

Another fictitious character was sent to me recently by a friend as a riddle. The character has three components:

女 上 下

The first component means "woman." The second and third components also are independent characters; they are pronounced *shàng* and *xià* and respectively mean "up" and "down." We assigned the signific to the "woman" component, and the character looked like this:

For the character as a whole we settled on the meaning "elevator girl." However, there is no intuitive way of pronouncing the character, since the last two components do not constitute a phonetic. The solution of pronouncing it with two syllables, *shàng-xià*, breaks a general rule of Chinese orthography, namely that for one character there is one syllable. Although reformers of the Chinese language have occasionally proposed polysyllabic characters, no such reform has ever been successful.

Chinese dictionaries and rhyme books may list tens of thousands of characters, but a knowledge of 4,000 to 7,000 characters is sufficient for, say, reading a newspaper. The form of the characters has been built up unsystematically through the centuries, and some are very intricate, requiring 30 or more strokes to write. The Chinese government's plan of simplification has gone a long way toward standardizing the form of the characters and reducing the average number of strokes per character. The net gain has been dramatic. A text written in simplified characters can contain fewer than half as many strokes as the same text written before 1956. This simplification makes the task of learning the written language considerably easier. An average of five or six strokes per character is not significantly different from the average of five or six letters per English word.

Since it is not the primary purpose of

the characters to represent sounds, the Chinese written language has been largely independent of the evolutionary changes that have taken place in the spoken language. This independence has made it possible for the written language to provide a literary continuity across thousands of years and to serve as a cohesive force binding the diverse cultures of China together.

CHARACTER FOR "SUN" is built up with four brushstrokes: first a vertical, then a turning stroke, then the inside and finally the closure. Order and the geometric position of the strokes are important. All the components of a character should fit roughly into a square.

The Evolution of the Language

The evolution of spoken Chinese, like the evolution of all other living languages, has been constant. Therefore many of the beautiful poems of the Táng dynasty of the seventh to 10th centuries no longer rhyme. If Confucius, who lived in the fifth century B.C., were to give a lecture anywhere in China today, he would not be understood. Within the large area of China dialects have evolved so far apart in their sounds that a man from Peking cannot be sure of being able to order a dinner in a Cantonese restaurant. Compared with the change in sounds, the written characters have changed little. Most of the characters Confucius used are still in books today, and many of these characters have their original meanings. The writing of Confucius is more intelligible to a modern Chinese than, say, a page of *Beowulf* is to an American. By the same token, although the Táng poems no longer rhyme, they are still enjoyed throughout China because their visual message remains the same. When the poems are read aloud by people in Peking, Shanghai or Canton, the poems sound altogether different because of the various dialects. It is rather like hearing "6 + 7 = 13" being read aloud in English, German and Norwegian. Even in Japan a Chinese with no knowledge of Japanese can manage to communicate reasonably well by writing. Chinese characters were also a significant medium of communication in Korea and in Vietnam. The independence of the characters from the spoken language has enabled them to serve as a core of culture in much of East Asia for many centuries.

A written Chinese character has a more direct connection with its meaning than a written word in English does. The sequence of letters spelling "horse" has meaning only through the mediation of the sounds they represent. The shape of the letters has no relation to the concept "horse." Little would be changed if English-speaking peoples were to take up the Cyrillic alphabet and the sounds for "horse" were represented *xopc*. To a Chi-

	OLD	SIMPLIFIED
SUN (rì)	日	日
STAR (xīng)	星	星
MORNING SUN (lŏng)	曨	昽
HORSE (mă)	馬	马
MOTHER (mā)	媽	妈
AGATE (mă)	瑪	玛
ANT (mă)	螞	蚂
TO SCOLD (mà)	罵	骂

SIMPLIFIED CHARACTERS were introduced in 1956 by the government of the People's Republic of China. Simple characters, such as the first two, were not affected. The remaining characters, somewhat more complex, were each reduced by six strokes.

nese the character for "horse" means horse with no mediation through the sound *mă*. The image is so vivid that one can almost sense an abstract figure galloping across the page:

马

The other major linguistic decision made by the Chinese government was to adopt a spelling system based on the Latin alphabet. This system is called Pinyin, which literally means "spell sound." All the Chinese words spelled out in Latin letters in this article are written in Pinyin. The government has been careful to point out that Pinyin is not intended to replace the characters but rather to serve as an aid in learning pronunciation. To discontinue the use of the Chinese characters would deprive coming generations of Chinese of a rich and meaningful cultural heritage.

Standard Chinese

To discuss the Pinyin system of spelling we need to examine the sound system of what is called standard Chinese, which is based on the Peking dialect and is now being taught in all parts of China. It is a straightforward system, simpler in many ways than the easy syllables of Spanish. There is really only one aspect unfamiliar to those who speak European languages: the tones. The musical quality of the spoken Chinese language is due to the fact that almost every syllable must carry one of four basic tones. These tones are indicated by diacritical marks over the vowels. The phenomenon of tones seems to be confined to Chinese and to some of the languages of Southeast Asia that have been heavily influenced by Chinese.

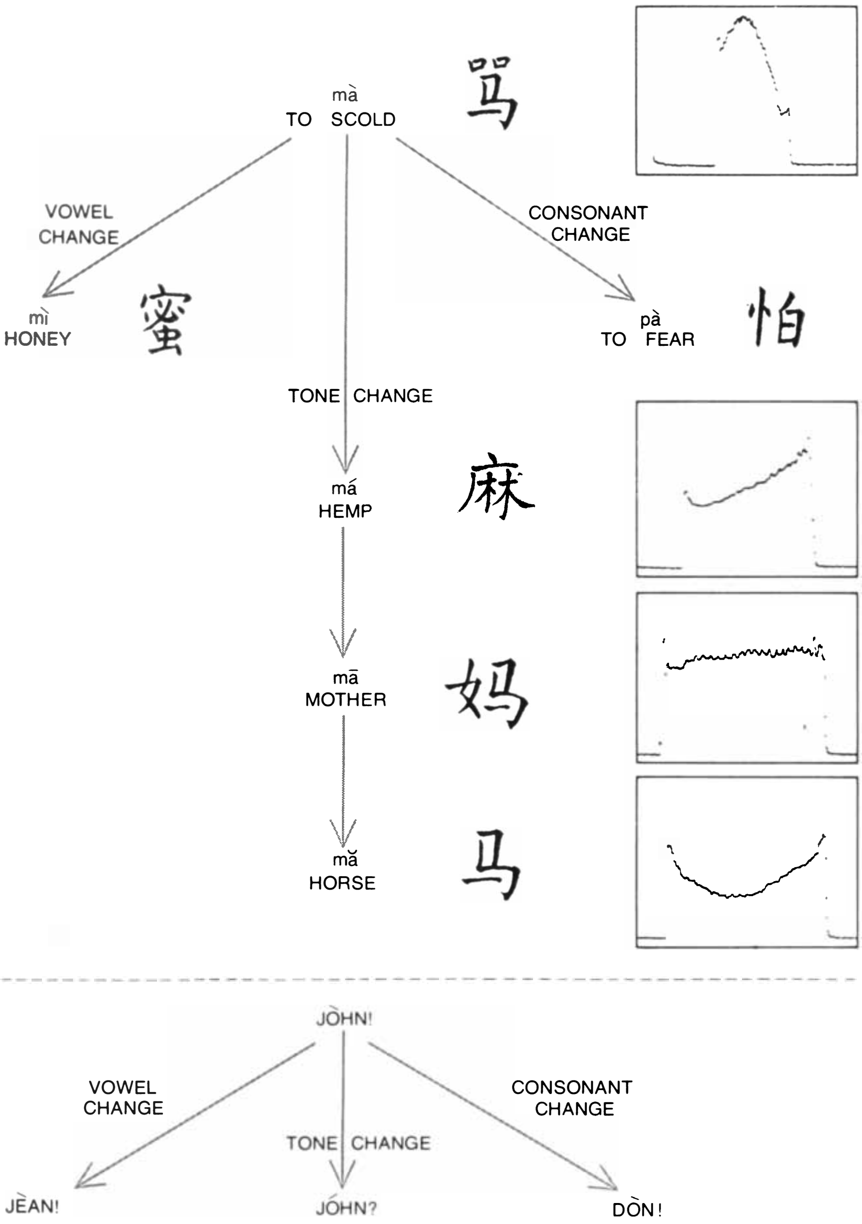
One of the earliest references in the literature to tones dates back to the sixth century. When the emperor of Liáng asked one of his scholars, Zhōu Shĕ, what was meant by the four tones, Zhōu responded with an elegant illustration:

天子圣哲

It means "The son of heaven is divine and wise." (Chinese emperors have traditionally been regarded as sons of heaven and divine, whether or not they were wise.) The beauty of Zhōu's response lies in the fact that the first word of his phrase illustrates the first tone in the speech of that time, the second word the second tone, and so on.

Essentially every Chinese syllable has a characteristic pitch pattern (tone). Changing the tone alters the meaning just as much as changing a consonant or a vowel in English changes the meaning. In English we use a rising pitch pattern for "Jóhn?" and a falling pitch pattern for "Jòhn!" The different tones convey different attitudes, but the meaning of the word remains the same. In Chinese, however, *má* with a rising pitch pattern means "hemp" and *mà* with a falling pitch pattern means "to scold." The two meanings are no more related to each other than they would be if we were to change the vowel to get *mì*, which means "honey," or if we were to change the consonant to get *pà*, which means "to fear."

Standard Chinese has a total of four tones: rising, falling, level and dipping [*see illustration at left*]. In addition to the tone every syllable must also have a nucleus to carry the tone, usually a vowel. The tone and the nucleus are the two obligatory components of the Chinese syllable. There are also three optional components of a syllable: the initial component, which is usually a consonant; the medial component, which is a glide, and the ending, which may be either a glide or a consonant from a restricted class. There are eight possible forms a syllable can take.



TONES are used to alter the meaning of Chinese words. Standard Chinese has only four tones: falling (as in *mà*), rising (*má*), level (*mā*) and dipping, or falling and then rising (*mă*). The oscillograph traces at right show the fundamental frequency of the author's voice as he spoke the words. In English, on the other hand, variation in tone is used to convey different moods; the meaning of the word being spoken does not change. In Chinese changing tone has same kind of effect on meaning of word as changing a vowel or a consonant.

			INITIAL CONSONANT	FINAL		
	PINYIN	PHONETIC		MEDIAL, GLIDE	NUCLEUS, VOWEL	ENDING, VOWEL OR CONSONANT
FACE	liǎn	ljɛ̌n	●	●	●	●
IRON	tiě	tʰjɛ̌	●	●	●	
TO GROW	zhǎng	tʂǎŋ	●		●	●
HORSE	mǎ	mǎ	●		●	
EYE	yǎn	jɛ̌n		●	●	●
MOON	yuè	yɛ̀		●	●	
HIDDEN	yǐn	ǐn			●	●
CHAIR	yǐ	ǐ			●	

COMPONENTS OF A CHINESE SYLLABLE that are obligatory are the tone and the nucleus (usually a vowel) to carry the tone. There are three optional components in a syllable: the initial, which is usually a consonant; the medial, which is a glide, and the ending, which may be either a glide or a consonant. All together there are eight possible forms that a Chinese syllable can take.

One striking feature of Chinese words in comparison with most European words is the lack of clusters of consonants before and after the nuclear vowel. When Western words with consonant clusters are represented in Chinese, they are typically broken up so that each consonant has its own syllable. "Marx" is conventionally rendered:

马 克 思

It is pronounced *mǎ-kè-sī.* The first character is the one for "horse," which also happens to be a prevalent Chinese surname.

Although the Pinyin system and standard Chinese are taught in all parts of China, the languages of ethnic minorities are given full consideration. According to *The Nationalities in China,* a book published in Peking in 1961, there are about 30 million people belonging to minority groups. The minorities occupy about a third of the land of China, mostly in the west and northwest. Some of the groups are large: the Zhuàngs are close to eight million in number and the Uighurs about four million. The languages of some of these groups are related genetically to Chinese but belong to other linguistic families such as Altaic and Austroasiatic.

A 1956 report by the Chinese Academy of Sciences estimated that the total number of people in China who spoke one or another of the Chinese dialects was more than 500 million. Of the dialects, Mandarin has by far the most speakers: more than two-thirds of the total [*see illustration on page* 52]. The Western name "Mandarin" derives from the fact that the dialect was the speech of government officials, or mandarins. It corresponds to the Chinese term *Guānhuà,* which means "official speech." Standard Chinese is a conventionalized variety of this dialect and is known as *Guóyŭ,* or "national language." It also is referred to as *Bĕifānghuà,* which means "northern speech," or as *Pŭtōnghuà,* "common speech."

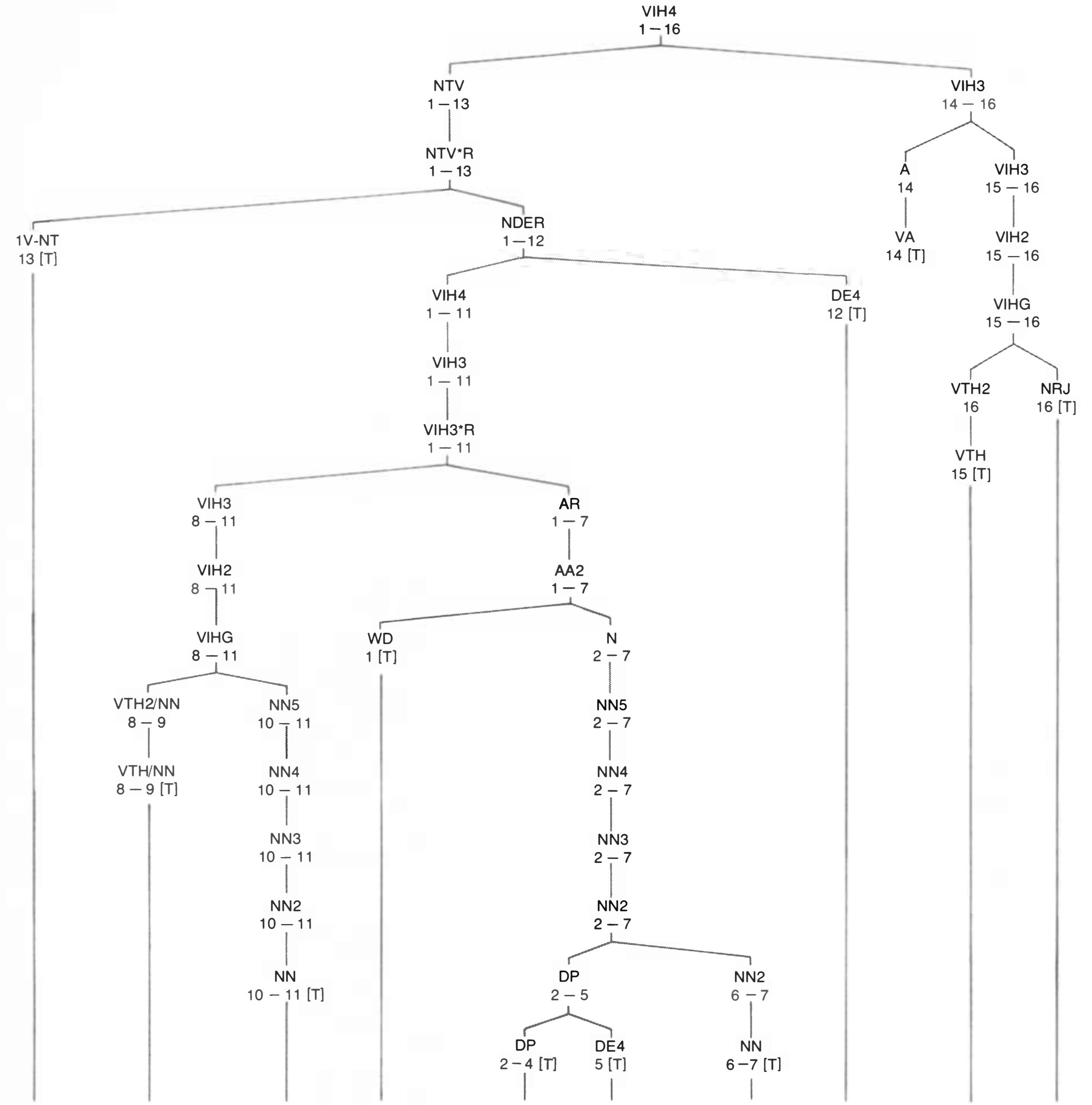
Outside China standard Chinese plays a much less important role. The dominant dialects among the 15 million Chinese in Southeast Asia are Yue and Southern Min, because it was people from the southern coastal provinces who migrated into the area. In the U.S. the Chinese are relative newcomers; they date back to 1850, when they were first drawn across the Pacific Ocean by the discovery of gold in California. Later these immigrants made up the core of the labor force that built the railroads in the American West. The ancestry of the great majority of the several hundred thousand Chinese now in the U.S. can be traced back to a small cluster of villages around Canton, all within a radius of 100 miles or so. The speech patterns of these villages are definitely of the Yue dialect group, but they differ markedly from one another. We can be sure that significant differences have arisen in the past 100 years between the speech of the American Chinese and the speech of the source villages around Canton.

Word Formation

Every language has a stock of several thousand morphemes: the bearers of the basic semantic and grammatical content. An expression such as "can openers" comprises four morphemes: "can," "open," "-er" and "-s." Some morphemes in English have more than one syllable (as in "open" and "adamant"), whereas others, such as the "-s" in "cats" and the "-t" in "slept," are single consonants. In Chinese most morphemes are exactly one syllable long. The usual division of morphemes into three major categories—noun, verb and adjective—applies to Chinese quite well. Thus in Chinese *shū* ("book") is a noun, *mǎi* ("buy") is a verb and *guì* ("expensive") is an adjective. These morphemes are known as contentives, in that they carry independent semantic meanings.

There is also a class of morphemes called functives. They are usually attached to contentives to modify their meaning in systematic ways and to show the relations the contentives have to one another. For example, functives attached to the contentive "prove" change its meaning: "proves," "proved," "proving," "disprove," "unproved" and "proof." Although conjugational and declensional paradigms are important in English, they are negligible in Chinese. There is, however, a sizable amount of derivational morphology in Chinese, where nouns are derived from verbs by changing the tone. Thus *shŭ* is a verb meaning "to count," and *shù* is a noun meaning "number." The verb *lián* means "to connect," and the noun *liàn* means "a chain." *Mó* is "to grind," and *mò* is "a grindstone." To derive a noun in these cases one simply changes the syllable to a falling tone. A similar example from English is deriving a noun from a verb by devoicing the final consonant: "prove"–"proof," "bathe"–"bath," "house"–"house."

The device of derivation by tone change is no longer productive in standard Chinese, but it is suspected that tonal derivation was an important process in the earlier stages of development



1V-NT	VTH/NN	NN	WD	DP	DE4	NN	DE4	VA	VTH	NRJ
13	8 – 9	10 – 11	1	2 – 4	5	6 – 7	12	14	15	16
1775	58972345	03374790	0110	755951746852	0037	63477820	0037	0668	1779	0037
HOU.4	CHONG.1 JI.1	YUAN.2 SU.4	YI.3	GAD.1 NENG.2 LIANG.4	ZHI.1	ZHI.2 DIAN.3	ZHI.1	KE.3	DE.3	ZHI.1
AFTER	[PHYS] BOMBARD	ELEMENT	WITH	HIGH-ENERGY		PARTICLE		MAY	OBTAIN	IT*THEM

以 高 能 量 之 質 点 衝 擊 元 素 之 後 可 得 之

INPUT STRING	0110	7559	5174	6852	0037	6347	7820	5897	2345	0337	4790	0037	1775	0668	1779	0037
SENTENCE POSITION	001	002	003	004	005	006	007	008	009	010	011	012	013	014	015	016

COMPUTER TRANSLATION of a Chinese sentence from a scientific text produces a reasonably accurate and understandable result. The string of Chinese characters is entered into the computer using a numeric code for each character. The position of each character in the sentence is also entered. The computer searches its memory for the meaning of each character and then performs syntactic analysis of the sentence. In converting the sentence into English the computer makes permutations of the word order. The sentence means: "It may be obtained after bombarding the element with high-energy particles." Research into computer analysis of Chinese is being conducted by the author and his colleagues at the phonology laboratory of the University of California at Berkeley.

of the Chinese language. In the Cantonese dialect, however, the process is still very productive. The Cantonese use tone change for forming diminutives: "candy" from "sugar," "daughter" from "female" and so on.

A common derivational device in Chinese is reduplication. Applied to nouns, it carries the meaning "every." Hence although *rén* means "person," *rénren* means "every person," *tiān* means "day" and *tiāntian* means "every day." Applied to verbs, it adds a transitory meaning to the action: *kàn* means "to look," whereas *kànkan* means "to take a look," *zŏu* means "to walk" and *zóuzou* means "to take a walk." Adverbs can be derived from adjectives by reduplication and the addition of a *de* suffix. Thus *kuài* is "quick" and *kuàikuàide* is "quickly," *lăn* is "lazy" and *lánlănde* is "lazily."

The manner in which two-syllable adjectives reduplicate is different. Whereas a two-syllable verb, *tăolùn* ("to discuss"), reduplicates as *tăolùntăolùn* ("to discuss a little"), an adjective, say *gāoxìng* ("happy"), becomes *gāogāoxìngxìngde* ("happily"). A verb reduplicates by the entire word, but the adjective reduplicates in terms of its constituent syllables.

Another device for word formation in Chinese is the conjoining of antonyms. "Buy" and "sell" combine to form "business," *măimài*. "Long" and "short" combine to form "length," *chángduăn*. The derived meaning is not always straightforward. For example, *făn* means "turned over" and *zhèng* means "right side up." Put together, *fănzhèng* means "in any case."

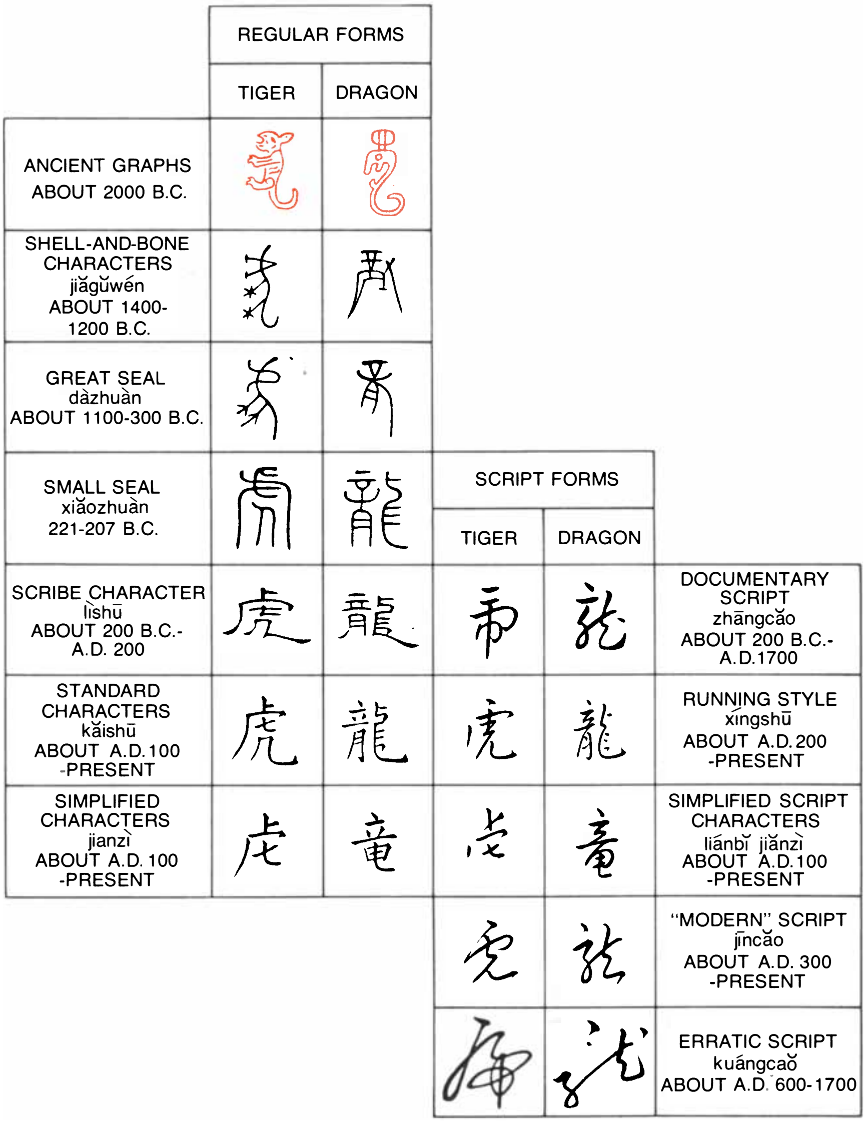
Classifiers are a linguistic feature peculiar to Chinese and its neighboring languages. Articles, numerals and other such modifiers cannot directly precede their associated noun; there has to be an intervening classifier, which usually has negligible semantic content. In Chinese one cannot say *sān shū* ("three books") or *nèi māo* ("that cat"). One has to say *sān bĕn shū* ("three piece book") or *nèi zhī māo* ("that piece cat"). The terms *bĕn* and *zhī* here are translated as "piece" for the lack of a better counterpart in English. Such classifiers are absolutely necessary in Chinese expressions. This feature has been carried over into many pidgin and Creole languages based on Chinese, in which "three bananas" or "this man" are rendered as "three piece banana" or "this fellow man."

Sentence Formation

The basic sentence in Chinese has the order subject-verb-object, as in English. Thus the sentence *wŏmen chī jī* is word for word "We eat chicken." There is a tendency in Chinese to delete either the subject or the object. Hence *wŏmen chī* ("We eat") or *chī jī* ("Eat chicken") are both common sentences. To focus attention on the object, the Chinese speaker will move it to the beginning of the sentence. *Jī wŏmen chī,* for instance, means "We eat chicken," but it is normally used to contrast with something else we do not eat. Suppose subject deletion and moving the object to the initial position both operated on the sentence *wŏmen chī jī*. First we front the object and get *jī wŏmen chī*, and then we delete the subject and get *jī chī*. The resulting sentence would mean something like "Chicken *A* eats," and the identity of *A* normally would be clear from the context. But *jī chī* of course also means something like "The chicken eats," where *jī* is the subject of the verb. In other words, *jī chī* is an ambiguous sentence, its ambiguity arising from object fronting and subject deletion.

There is no evidence that Chinese allows either more ambiguity than English or less. In an English sentence such as "It is too hot to eat," the "it" can refer to the weather, to the food or to the animal that is doing the eating. Moreover, "hot" could mean "high in heat content" (which in Chinese is *tàng*) or it could mean "spicy" (which in Chinese is *là*).

The Chinese language as an object of study goes back as far as the beginning



HISTORICAL DEVELOPMENT of pictographic characters for the two most powerful animals in Chinese mythology, the tiger and the dragon, are shown in their various stages. The earliest-known pictographic forms of the animals are at the top in red. The illustration is from *Introduction to Chinese Cursive Script*, by F. Y. Wang of Seton Hall University.

of the Christian Era. At that time the Chinese had already produced sophisticated works in dialectology and in semantic classification. The foundations for the historical study of Chinese sounds were laid during the 17th and 18th centuries by the great scholars of the Qīng dynasty, and it is on their shoulders that modern Chinese linguists stand.

The key to how a language sounded centuries ago lies in how it sounds today. The basic method is to compare the pronunciation of morphemes in contemporary dialects and to infer what their ancestral pronunciations might have been. The inference is not just a guess but is made on the basis of documentary evidence and knowledge of the general linguistic principles underlying sound change.

Given the nonphonetic nature of the Chinese writing system, it may seem an impossible task to reconstruct how the language was spoken many centuries ago. One might think that the sounds of alphabetic languages, such as Old Church Slavic or Sanskrit, would be much easier to reconstruct. Actually it is not much easier, because there is no direct way to determine how a letter was pronounced. With alphabetic languages the phonetic values must also be arrived at by inference.

Chinese has the great advantage of an abundance of ancient writings that reach back continuously in time further than the literature of any other language in the world. The fact that the form of Chinese characters is often not much influenced by changes in pronunciation is quite a convenience in helping linguists to determine which morphemes are etymologically related.

The study of Chinese dialects has been hampered in the past by an overabundance of data, which tended to make research procedures cumbersome and time-consuming. The advent of large computers has facilitated the manipulation of the data. Chin-Chuan Cheng and I, with the help of several colleagues, have developed a dialect dictionary on computer, which we call DOC. The program is in operation at the computer centers of the University of California at Berkeley and of the University of Illinois. It incorporates the pronunciation of more than 2,000 morphemes in each of 20 Chinese dialects.

Evidence is accumulating from our work with DOC that changes in language proceed in ways that are essentially parallel to biological evolution, as Charles Darwin noted in *The Descent of Man.* In both cases the mechanism of change resides in variation. When two or more variants appear, the rival forms compete for survival. For instance, in American speech the vowel in "room" varies between that of "pool" and that of "put," and the *s* in "disobey" varies between an *s* and a *z*. The major selective force, which is constant across time as well as across languages, is the ease with which the forms can be pronounced and properly perceived. This selective force determines what forms of speech will survive.

With the aid of computer programs such as DOC we have been able to run through large pools of data to locate ongoing sound changes of special interest. Some changes are just beginning, some are in midstream and some are ending. In the Cháozhōu dialect, for example, there is a change from one tone to another that has so far affected about half of the vocabulary. By having access to a large number of well-defined types of sound change, we are now in a position to study much more effectively the selective force that determines the direction of language evolution.

Another aspect of the Chinese language that has been subjected to intensive computer study is its morphology and syntax. With the collaboration of Stephen W. Chan, Benjamin K. Tsou and others, we have developed a machine dictionary with more than 70,000 entries, together with the necessary programs for translation from Chinese into English. A good measure of how well we understand the structure of a language is how well we can break it apart in a way that is suitable for translation by machine. Although we are an indefinite distance away from being able to translate a Táng poem into English without losing its exquisite sensitivity, we can do a reasonably accurate job with scientific texts [*see illustration on page* 58].

Now that relations between the People's Republic of China and Western nations are becoming more normal, interest in the Chinese language is increasing at an accelerating tempo. Considering that the Chinese language has the largest number of speakers in the world and the greatest time depth in its literature, this interest is long overdue. With the increase in interest we may look forward to deeper probings into the history and the structure of the language, and into the influence the language has had on the cultural and intellectual development of the Chinese people. These studies will surely lead to a better general understanding of the nature of human language.

Famous 3-Way Doorgate.

New recreation table.

Load space takes 4 x 8 building panels.

Reclining passenger seat.

Electric rear window defroster.

Power Mini-Vent windows.

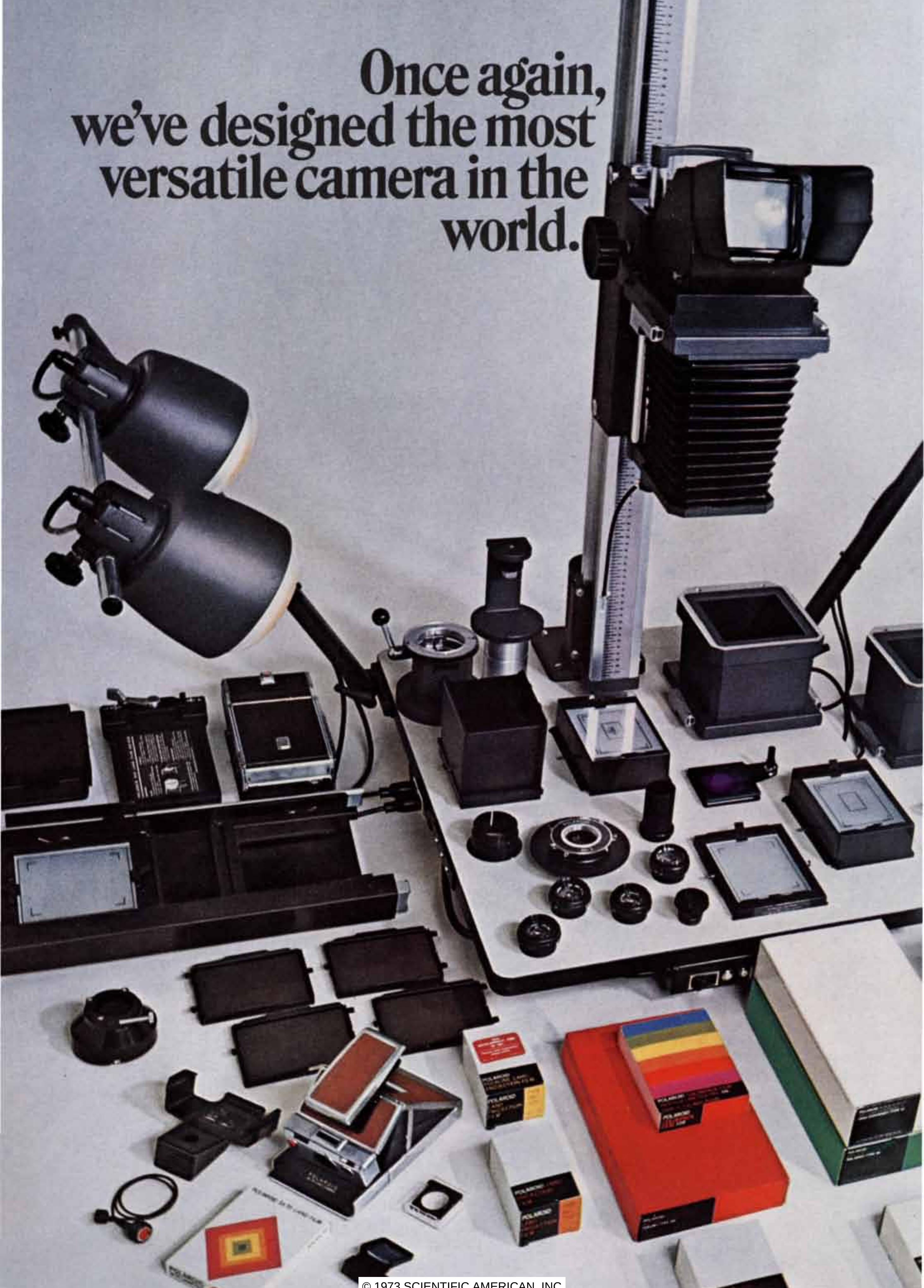
One reason more people buy Ford wagons is our wide variety of features. Like the 3-Way Doorgate that opens out for people and down for cargo, and Ford's dual facing rear seats. Now we've added windshield washer jets mounted on the wiper arms. A spare tire extractor. Optional Power Mini-Vent windows. Ford wagons include the full-size Country Squire (above), mid-size Torinos and little Pintos. All with energy-absorbing bumpers and optional steel-belted radial ply tires tested to give the average driver 40,000 miles of tread life with normal driving. At your Ford Dealer.

LTD Country Squire. Standard features include woodgrain bodyside, tailgate treatment, 3-Way Doorgate, automatic transmission, power steering and brakes. Other equipment shown is optional.

When it comes to wagons, nobody swings like Ford.

FORD WAGONS

FORD DIVISION Ford


Once again,
we've designed the most
versatile camera in the
world.

Once we'd designed our first multipurpose camera, we probably could have kept on selling it forever. Because it was the most versatile camera in the world.

But we're obsessive perfectionists, so we couldn't keep ourselves from improving our original. And we improved it in so many ways (16 to be precise) that it became a whole new camera: The Polaroid MP-4 Multipurpose Land Camera.

The MP-4 may *look* familiar. (And some of the things it does may *be* familiar. Like delivering color pictures in a minute, and black-and-white in seconds.) But it has a new range of capabilities.

For example, its modular design, coupled with a wider choice of accessories, lets you build (and pay for) only the exact photographic system you need.

If you want additional capabilities later on, you can add them later on.

It's easy to get optimal lighting for your subjects because the lighting arms are now completely adjustable. And the lamp heads, themselves, are removable (with heat shields) so they can be shifted anywhere. Even the switches and accessory outlets are more conveniently located.

Macrophotographs (formerly up to 10 times actual size) can now be made up to 14.8 times actual size. And critical focusing for high macro- and photomicroshots is facilitated by the addition of an aerial image ground glass.

With the MP-4, subjects can be subjected to photographic scrutiny from almost any angle without being repositioned. Furthermore, when the unit is on a table, it can shoot charts on the wall or large objects on the floor. These feats are possible because the MP-4 column can rotate 360° around its vertical axis, while the camera can rotate 360° around its horizontal axis.

But the MP-4 camera is more than just movable. It is *removable* from the column. So when the subject can't come to the camera, the camera can go to the subject. At which point the camera with its tripod mount becomes a studio camera.

We could go on about the new MP-4 enlarger head, the self-cocking Copal shutter that accepts 5 Tominon lenses or the universal mount that lets you use other cameras *including* Polaroid's SX-70* (a camera that revolutionizes instant photography the way the first Land camera revolutionized conventional photography).

But you get the point. We took an incredibly versatile camera and made it even more versatile in many ways. And more sensible in other ways. And more economical in some ways.

If you want a demonstration or more information, write to Polaroid Corporation, Dept. 57-257, Cambridge, Mass. 02139. (In Canada: 350 Carlingview Dr., Rexdale, Ont. In Puerto Rico: P.O. Box 2032, Ceramica Annex, Carolina 00630.)

WHAT IT CAN DO ON THE SPOT.

Copy wall charts
Enlarge negatives
Make 4 x 5 prints
Copy graphs
Record test results
Make 3¼ x 4¼ prints
Copy EKG's
Copy 35mm slides
Make 3¼ x 4 slides
Record fingerprints
Copy X-rays
Make product pictures
Make metallographs
Record defective parts
Photograph models
Make halftone prints
Record gross specimens
Illustrate research reports
Record legal evidence
Make close-ups of samples
Copy original artwork
Size type for mechanicals
Record electrophoresis results
Document research projects
Make positive/negative photographs
Make photomicrographs
Create graphic designs
Produce catalog artwork
Make macrophotographs
Take "before" and "after" records
Make copyprints of original photos
Illustrate training procedures
Record destructive tests
Copy engineering drawings
Record assembly sequences
Convert continuous tone material to line art
Copy results of nondestructive tests
Photograph transient phenomena
Photograph fluorescent phenomena
Make record of prototypes
Record insured valuables
Produce advertising pictures
Record bacterial cultures
Copy medical illustrations
"Memorize" laboratory setups
Copy scintillation photos
Photograph chromosomes for karyotyping
Et cetera.

The new Polaroid MP-4 Multipurpose Land Camera.

*Available only in the U.S. at this time. Polaroid®

THE MICROSTRUCTURE OF THE OCEAN

The temperature and saltiness of seawater can now be mapped centimeter by centimeter. Such mapping is needed to learn how the sea is so effectively stirred by the winds and the tides

by Michael C. Gregg

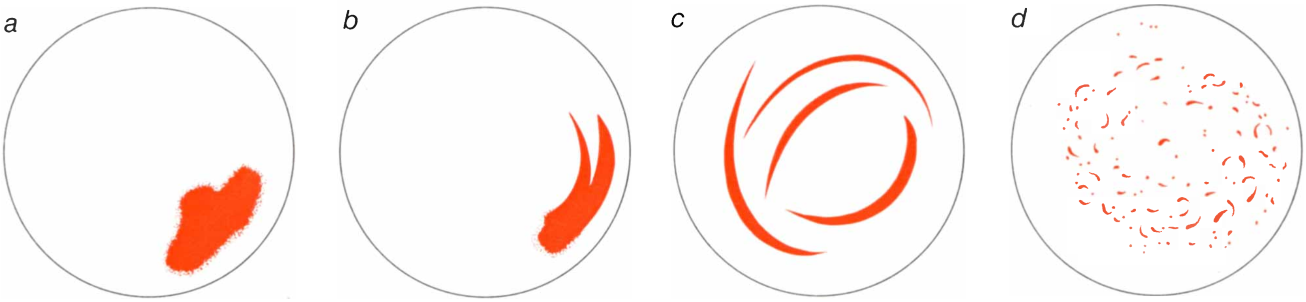
It is just 100 years and two months since H.M.S. *Challenger,* a wooden corvette of 2,306 tons, sailed from Portsmouth, England, on the world's first oceanographic expedition. During the voyage, which lasted four years and covered 69,000 nautical miles, the *Challenger* investigators made observations at 362 locations. As a part of their mission they took tens of thousands of samples of the Atlantic, Pacific and Antarctic oceans by lowering bottles that could be opened at a prescribed depth, filled and sealed before being hoisted back aboard ship. At each depth a temperature reading was taken by breaking the capillary column in a bulb thermometer by remote control. This cumbersome method of collecting samples and taking temperatures prevailed with modest improvements for the next 90 years and has provided much of what we know about the large-scale characteristics of the medium that covers 70 percent of the earth's surface. Although the sample-bottle-and-thermometer method is a good one for establishing the gross seasonal changes in salinity, temperature and density of seawater at various depths, its sampling "mesh" is far too coarse to detect the small-scale transactions in salinity, temperature and density by which energy received from the sun is exchanged between the ocean and the atmosphere and ultimately redistributed over the earth.

Within the past 10 years advances in solid-state electronics have made possible rugged instruments that can supply a continuous record of temperature, salinity, density and velocity from the ocean surface down to depths of a kilometer or more. Because these new instruments are able to sense changes over distances of less than a centimeter they have revealed the widespread existence of coherent structures in the sea that previously escaped detection. In some places steplike profiles of temperature and salinity suggest the presence of homogeneous layers of water separated by thin interfaces. Typical layers 10 to 20 meters thick can often be traced horizontally for as far as 15 kilometers; other layers only 20 to 30 centimeters thick extend laterally for several hundred meters. Vertical measurements in the range from a few millimeters to a few centimeters reveal intense and irregular fluctuations of temperature and salinity, indicating that the top 30 meters of the sea is often completely turbulent. It is here that energy is being exchanged most actively. It is now customary to apply the term "microstructure" to oceanic physical processes on the scale of a few centimeters or less. As we shall see, the rate of diffusion of temperature and salinity along microstructure gradients sets a limit to the smallest significant scales of fluctuation in the ocean.

It has been evident since the voyage of the *Challenger* that the gross features of the principal ocean-current systems and the distribution patterns of temperature and salinity change little from year to year; they represent nearly a steady-state condition. By learning how processes on the microscale lead to this steady-state condition of temperature and salinity we shall be better able to understand how substances such as nutrients and man-made pollutants are distributed in the ocean.

Two closely related processes are responsible for dissipating energy and smoothing variations in the ocean: mixing and stirring. Although the two terms are often used interchangeably in ordinary speech, the oceanographer must make an important distinction between them. This distinction was pointed out by Carl Eckart of the Scripps Institution of Oceanography in a farseeing paper published in 1948. Mixing is the reduction of variations by the action of molecular diffusion along gradients from regions of higher temperature or concentration to regions of lower temperature or concentration. The rate of mixing depends directly on the steepness of the gradients and on the molecular coefficients of diffusivity. Stirring, on the other hand, is the creation of velocity differences by any process that imparts kinetic energy to a liquid. If a liquid is not homogeneous with respect to, say, temperature and salinity, stirring will tend to increase the gradients between nonhomogeneous parcels, leading both to greater variability and to faster mixing. For example, if a source of variations is introduced by carefully injecting a blob of salty water into an otherwise uniform and still sample of fresh water, the blob will sink to the bottom and persist for a considerable time because salt diffuses rather slowly. If one stirs the water with a paddle, the resulting velocity differences, or shear, will distort and stretch the blob, thus increasing the area of contact between it and the surrounding water and sharpening the salinity gradi-

WAKE LEFT BY A DYE PELLET is photographed by a diver (*at left in the photograph on the opposite page*) in a study of oceanic microstructure conducted in the Mediterranean by John D. Woods of the University of Southampton. Kinks in the wake of the pellet reveal differences in the motion of adjacent layers in the water. The marker to the right of the dye-pellet wake is held in position by other divers to provide a scale and a reference to the vertical and horizontal. Divers working with Woods were members of a Royal Navy team.



PROCESS OF STIRRING involves the creation of velocity differences in a fluid by the introduction of kinetic energy. The term "mixing" refers to the processes of molecular diffusion that spontaneously reduce the variations in temperature or chemical concentration that exist in a nonhomogeneous system. If a blob of salty water (*color*) is gently introduced into a still body of fresh water (*a*), the blob will persist for a long time unless the mixing is hastened by stirring. If the paddle is moved through the water, it produces differences in velocity that stretch and distort the blob, increasing its area and sharpening the salinity gradients (*b*). If the stirring is now continued (*c*), the blob is broken into long, thin filaments. Both the increase in area and the sharpening of gradients increase the rate of molecular diffusion from regions of higher salinity to regions of lower salinity. With or without stirring, the mixing will eventually approach completion (*d*). The winds and the tides are the principal agencies that act to stir the ocean.

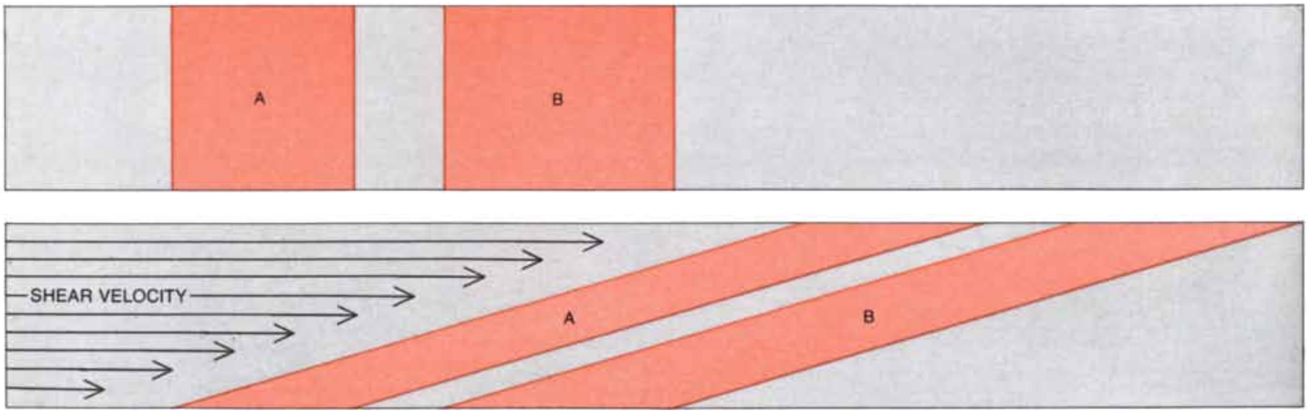
ents across the interface [*see illustration above*]. Both of these effects will hasten mixing.

An important task of the oceanographer is to identify the forces that stir the ocean. Stirring motions in the deep ocean are usually gentle. Although these weak shears produce marked contrasts in temperature and salinity over distances greater than several meters, they generate little microstructure and therefore do little to enhance the rate of mixing. A common situation arises when two adjacent volumes of water that differ significantly in temperature are sheared sideways by a vertical velocity gradient; the result is a strong temperature contrast in the vertical [*see illustration below*]. This mechanism probably accounts for many features in large-scale vertical temperature profiles and in some of the smaller "stairstep" profiles observed near the surface.

We have substantial evidence for two stirring processes whose strong shearing motions are capable of generating intense microscale activity. The first is turbulent stirring driven either by the winds near the surface or by shear instabilities associated with currents and internal waves at greater depths. The second stirring process arises from vertical convective motions produced by the differing rates of diffusion of heat and salt. This process is a local one whose vertical scale seldom exceeds a few meters; it tends to operate where weak stirring has already increased the vertical gradients of temperature and salinity. Before going further into the processes that generate microstructure I shall discuss how variations on a scale exceeding several meters arise. I shall deal primarily with vertical variations both because they are more pronounced than horizontal ones and because they are better understood.

Since the sea surface is nearly in thermal equilibrium with the atmosphere, variations in surface temperature reflect the local heat balance, which ranges from −1.9 degrees Celsius in polar regions to nearly 30 degrees C. at the Equator. Away from coastal areas, where the runoff of precipitation affects the salt content, the salinity of the ocean varies from 32 to 37 grams of dissolved solid per kilogram of seawater, expressed as 32 to 37 ‰. (The symbol ‰ is read "per mil.") The variations in salinity reflect differences in the relative rates of precipitation and of evaporation from the surface of the ocean. When differences in salinity and temperature produce horizontal differences in density, there is a tendency for the denser waters to sink below those that are less dense.

Although a difference of 1 ‰ in salinity has an effect on the density of seawater about five times greater than a difference of one degree C. in temperature, the greater range of temperature dominates the large-scale density structure of the ocean. Salinity differences, however, have very appreciable effects on smaller scale. The densest waters are believed to be formed under the ice of the Weddell Sea in the Antarctic, in the North Atlantic off Greenland and in the Norwegian Sea. These dense waters sink to the bot-



STRONG VERTICAL TEMPERATURE GRADIENTS can be formed when two adjacent volumes of water at different temperatures (*A*, *B*) are tilted sideways by shearing currents. Illustrations on this page follow diagrams first published by Carl Eckart.

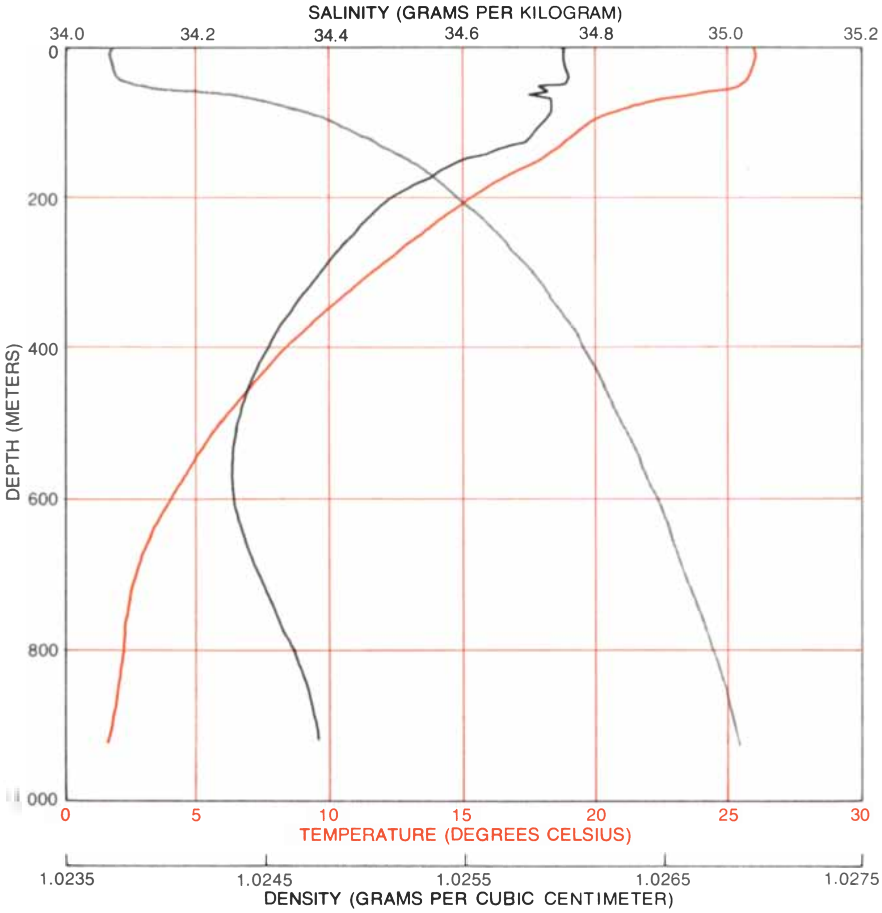
tom and slowly spread into the major ocean basins. Less dense waters from other high-latitude sites sink to intermediate depths in a similar manner. As they spread into the central oceans the various tongues of water interleave, forming vertical density profiles with a continuous stratification. Since any ocean is very shallow in relation to its breadth, the interleaving of water of various densities gives rise to temperature gradients that are much sharper vertically than horizontally.

The structure of the vertical temperature gradients is seen most simply in the open ocean, far from complicating boundary effects. Here measurements of temperature and salinity show little change from one location to another or from year to year at the same location. Below the well-mixed layer at the surface the temperature drops very rapidly. Below the region where the temperature gradient is a maximum, known as the thermocline, the temperature profile is very nearly an exponential curve [*see illustration at right*]. The temperature varies smoothly over vertical distances greater than several meters, implying that the tongues of water forming the temperature profile have undergone considerable vertical mixing as they spread toward the Equator from high latitudes. Many theoretical treatments, known as thermocline models, have attempted to establish how this mixing takes place. Such studies require that the amount of water sinking at high latitudes be balanced by the amount rising in low or middle latitudes.

If one assumes in a thermocline model that there is a broad general upward movement of the bottom water over the interior of the ocean basins, one must provide for a moderately intense downward transport of heat. Since molecular heat conduction (that is, simple mixing) along the mean temperature gradients is too slow to match the upward movement of the cold water, some models assume the existence of additional mixing driven by turbulent stirring to increase the heat flow [*see top illustration on page 70*]. Typical values obtained by combining the downward turbulence-driven heat flow and the upward motion of the cold deep water to fit the observed temperature data are an upward velocity of one centimeter per day and a turbulent-mixing coefficient of one square centimeter per second. It is assumed that the rate of mixing is uniform except within a few hundred meters of the surface and the bottom.

An alternative model suggests that enhanced mixing rates may not be necessary if the effects of prevailing winds are considered. The Coriolis force, which is created by the earth's rotation, tends to deflect moving objects to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. Early in this century the Norwegian oceanographer Vagn Walfrid Ekman concluded that the combined effect of a steady wind and the Coriolis force is to produce a net transport of the surface water 90 degrees to the right of the wind in the Northern Hemisphere and an equal amount to the left south of the Equator.

Thus the prevailing westerlies in the middle latitudes and the easterly trade winds in the Tropics combine to drive the surface waters to the center of subtropical gyres (large, slowly rotating masses of water). The net inflow along the surface leads to a slight elevation in the center of the gyres and a tendency for the warm waters under the gyres to sink [*see bottom illustration on page 70*]. The downward heat flux provides a mechanism for the formation of the thermocline without the necessity of invoking turbulence-driven mixing. In this model the rising deep waters cannot penetrate the thermocline over a broad region but are concentrated in restricted areas, such as at the Equator and near the coasts. The two types of model make very different predictions of the rates of mixing to be expected in the deep ocean.



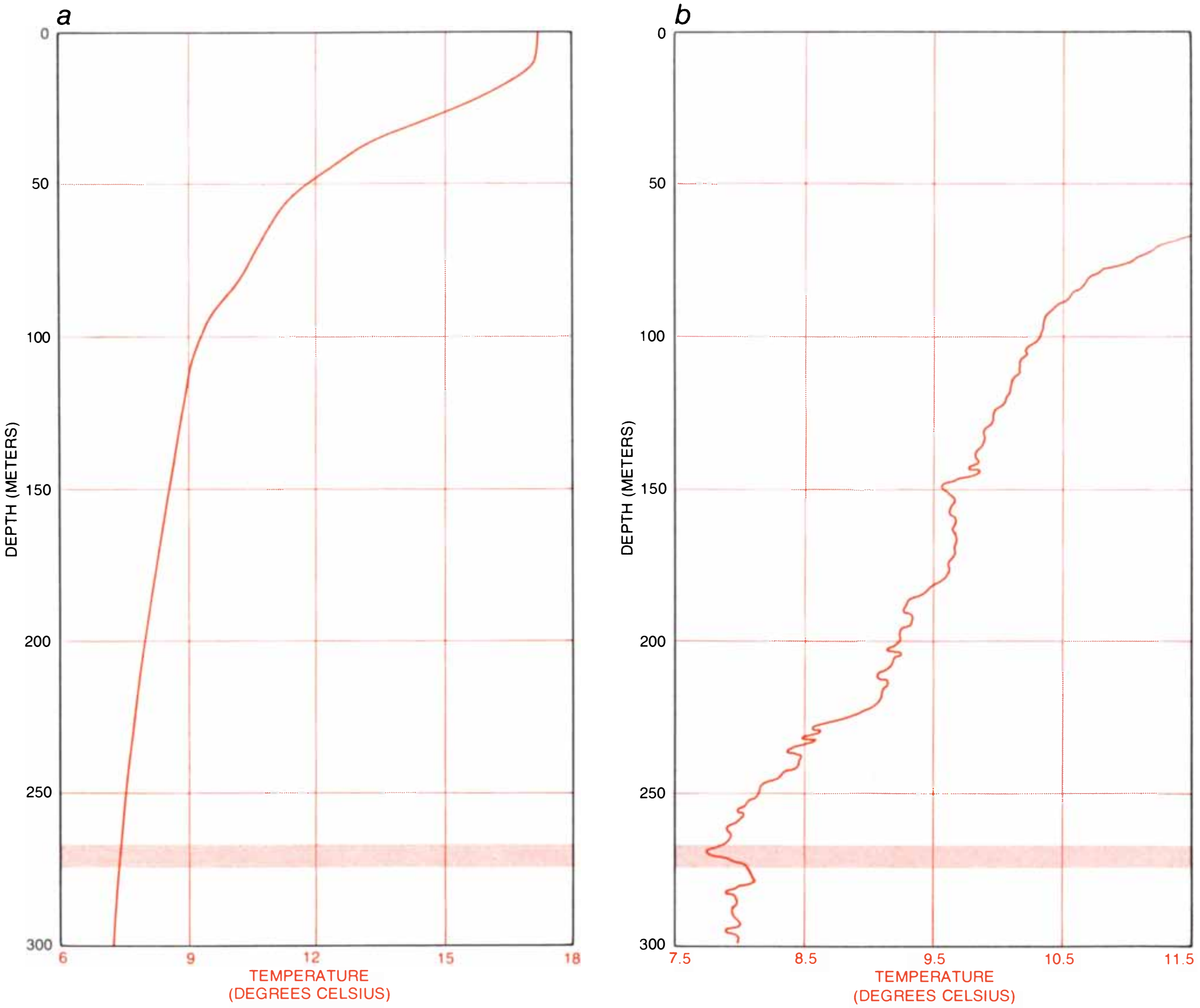
TEMPERATURE AND SALINITY IN MID-PACIFIC were recorded with an instrument having a resolution of a few meters. A well-mixed layer in which the temperature and salinity are nearly constant extends to a depth of about 60 meters. Their temperature (*colored curve*) falls steeply and then more slowly. The region of rapid temperature change extending several hundred meters below the mixed layer is called the thermocline. The salinity (*black curve*) reaches a minimum value at about 550 meters, then rises again. Since the temperature dominates the density structure (*gray curve*), the region of maximum density change coincides with the thermocline. The temperature of the Pacific below a depth of four kilometers is about one degree Celsius; salinity is 34.7 grams per kilogram of seawater.

Temperature variations on an intermediate scale, with vertical dimensions from a few meters to several hundred meters, are most pronounced in coastal regions and near the open-ocean boundaries of major current systems. The variations are due in part to variations in meteorological conditions, which can lead to the formation of local tongues of dense water, and to the stirring effect of the rather high shears associated with the surface and subsurface boundary currents. For example, vertical measurements made a few miles off San Diego show pronounced and irregular layers 10 to 30 meters thick, including numerous temperature inversions, that is, regions where the temperature increases with depth instead of decreasing [*see illustration on these two pages*].

These layers, a characteristic feature below the California Current, are believed to be formed by the interleaving of discordant water types of different densities. The temperature inversions are accompanied by local increases in salinity that compensate for the reduced density associated with the warmer layers of water, so that the net density increases continuously with depth. The warm saline water is formed in the eastern tropical Pacific and flows northward at depths of 200 to 300 meters along the California coast, where it intermingles with cold, less salty water that is flowing south along the surface from the Gulf of Alaska.

High-resolution observations made with continuous recorders have shown that below the near-surface regions high vertical mixing rates tend to be found in small patches where there is intense microstructure activity. The rate of molecular diffusion along these thin but sharp microstructure gradients is at least several factors of 10 higher than the diffusion rate found in the much weaker large-scale gradients. The evidence to date suggests that the patches of microstructure activity and high mixing rates have vertical dimensions of a few meters or less. Most of the gradient features of temperature and salinity profiles greater than 10 meters or so are therefore not the result of mixing but of weak shearing motions. These larger structures are continuously being modified and smoothed by the existence of smaller regions in which there is intense mixing.

Some degree of turbulence is associated with all these local regions of high mixing. In the dynamical instabilities, where motions tend to become disorganized, the role of turbulent stirring is dominant. In the differential diffusion processes, on the other hand, where convective instabilities lead to more orderly velocity patterns, the role of turbulence is more restricted. On scales of a few meters, however, the distinction between features formed by the processes that

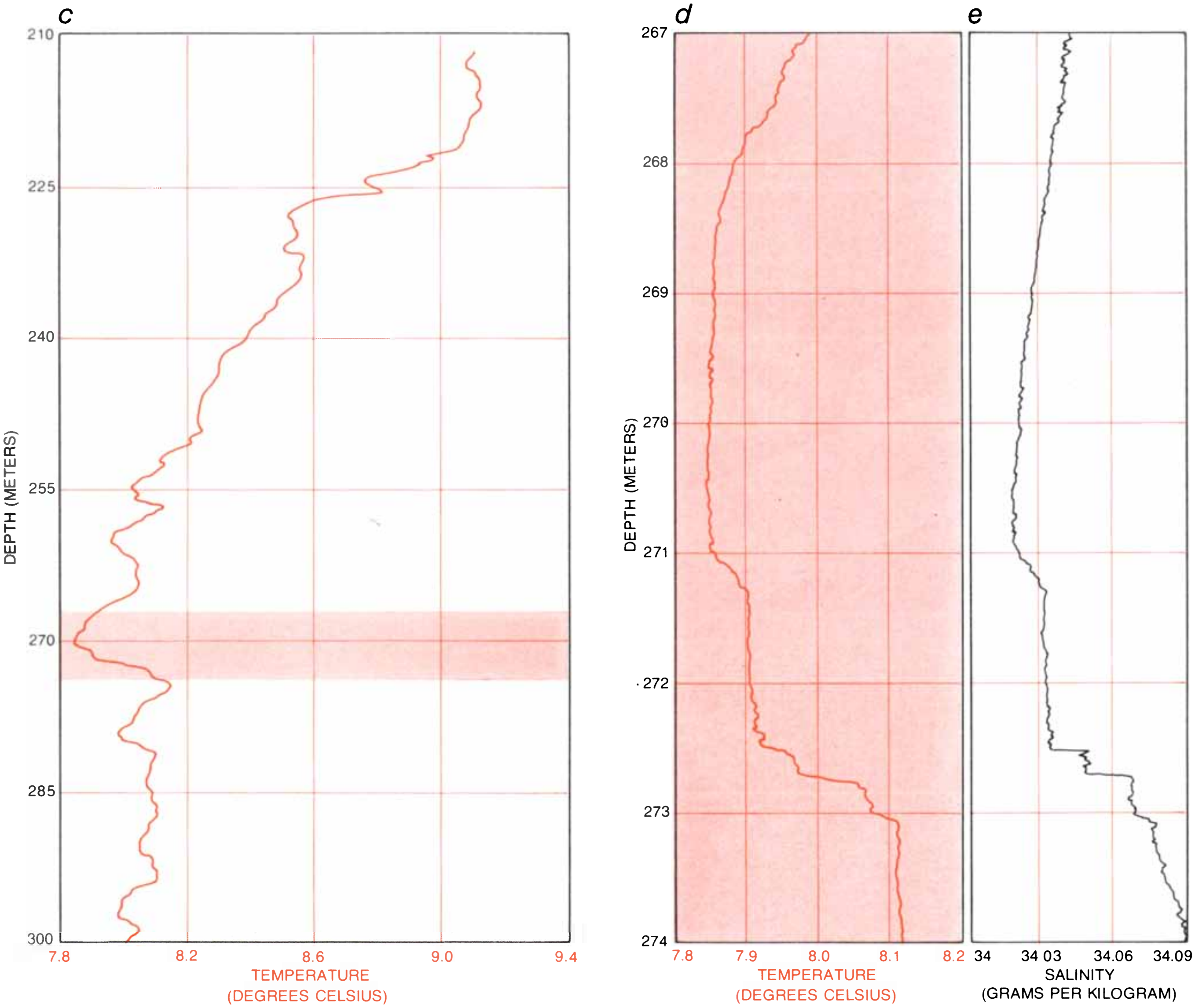


SERIES OF TEMPERATURE READINGS shows how small-scale variations that eluded detection by earlier methods are revealed by electronic recorders. The first curve (*a*) is a hand-plotted record of the ocean temperature off San Diego made by lowering a bulb thermometer, a simple method that dates back more than a century. The second curve (*b*) was made nearby with a commercial electronic instrument, which provides a continuous record with a maximum resolution of a few meters. The third trace (*c*) was made simultaneously with a still more sensitive instrument capable of revealing temperature fluctuations centimeter by centimeter. An generate microstructure and features due to weaker motions that do not result in much mixing is complicated by the fact that away from the surface the two kinds of process rarely carry the mixing to completion. In a stratified fluid such partial mixing usually results in the formation of additional layers.

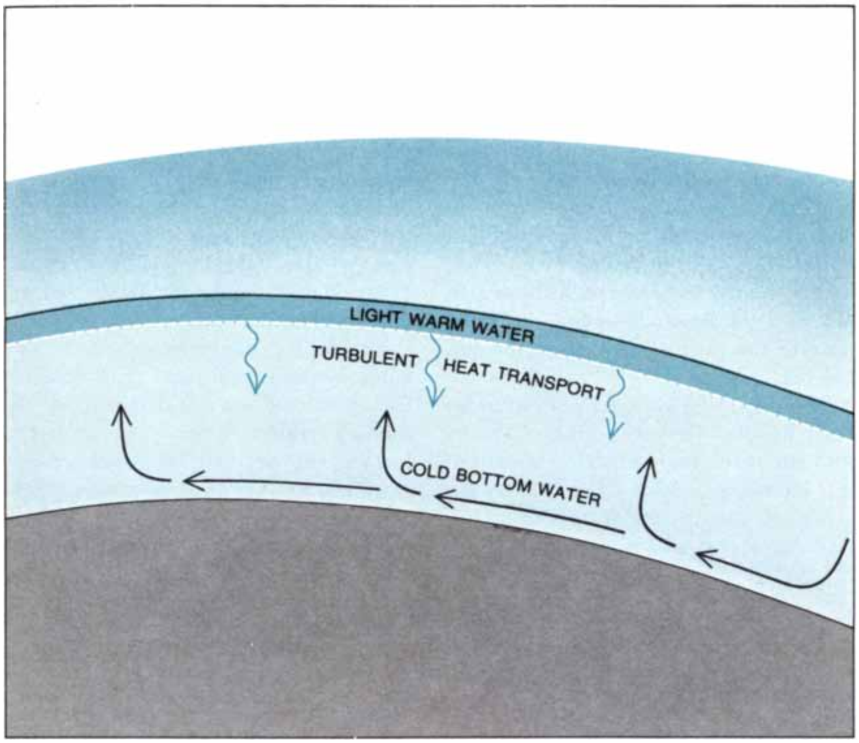
The task of providing a theoretical description of turbulence has been a challenge to generations of hydrodynamicists. Since the random velocity fluctuations that characterize turbulence continue to elude a rigorous theory, the most successful treatments have either been statistical or have relied on physical arguments and dimensional analysis. Briefly stated, dimensional analysis involves determining what parameters are present in an equation and then deducing how they must be related to obtain the same dimensions on both sides of the equation. For example, a velocity scale (in centimeters per second) must be the ratio of a length scale (in centimeters) and a time scale (in seconds). It has proved to be useful to think of turbulent flow as a superposition of eddies of different sizes, the largest of which are limited by the dimensions of the region in which the flow takes place or by the scale of the process generating the turbulence.

The energy extracted from the mean flow by the largest eddies cascades through increasingly smaller eddies until it is ultimately dissipated as heat by the action of viscosity. If the turbulence is well developed, so that the smallest eddies are many factors of 10 smaller than the large eddies that hold the bulk of the kinetic energy, it turns out that the motions of smallest scale are statistically independent of the largest eddies. The successive cascading of energy through increasingly smaller scales of motion decouples the smallest eddies from the details of the large-scale flow, so that the energy balance for the small-scale events depends only on the dynamics that are important at that scale.

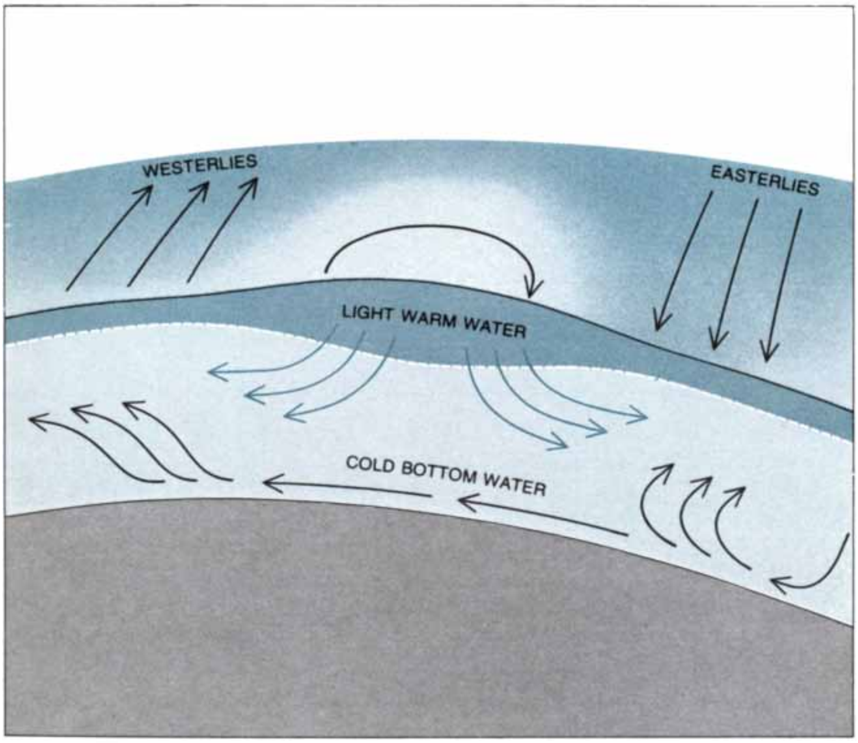
The Russian mathematician A. N. Kolmogorov postulated some years ago that the dominant energy balance of the smallest eddies depends on only two factors: the net rate at which energy cascades down from large-scale motions and the viscosity of the fluid. Kolmogorov showed by dimensional analysis



enlarged section of the third record, covering tiny temperature variations over a depth interval of only seven meters, constitutes the fourth trace (*d*). The seven-meter interval from 267 to 274 meters is indicated by a colored band in the first three traces. The last curve (*e*) is a salinity record made simultaneously with the high-resolution temperature record at its left. One can see a step-like layering of temperature and salinity between 271 and 274 meters. Such records provide clues to the transport of heat and chemical elements in the ocean. These records were made by the author and Charles S. Cox of the Scripps Institution of Oceanography.



THERMOCLINE MODELS attempt to explain how heat is exchanged between the warm surface waters of the subtropics and the cold deep water that originates in polar regions. The exchange produces the characteristic mid-ocean temperature profile that is shown in the illustration on page 67. This simple model assumes that the downward mixing of heat, driven by turbulent stirring, is balanced by a slow upward drift of cold bottom water.

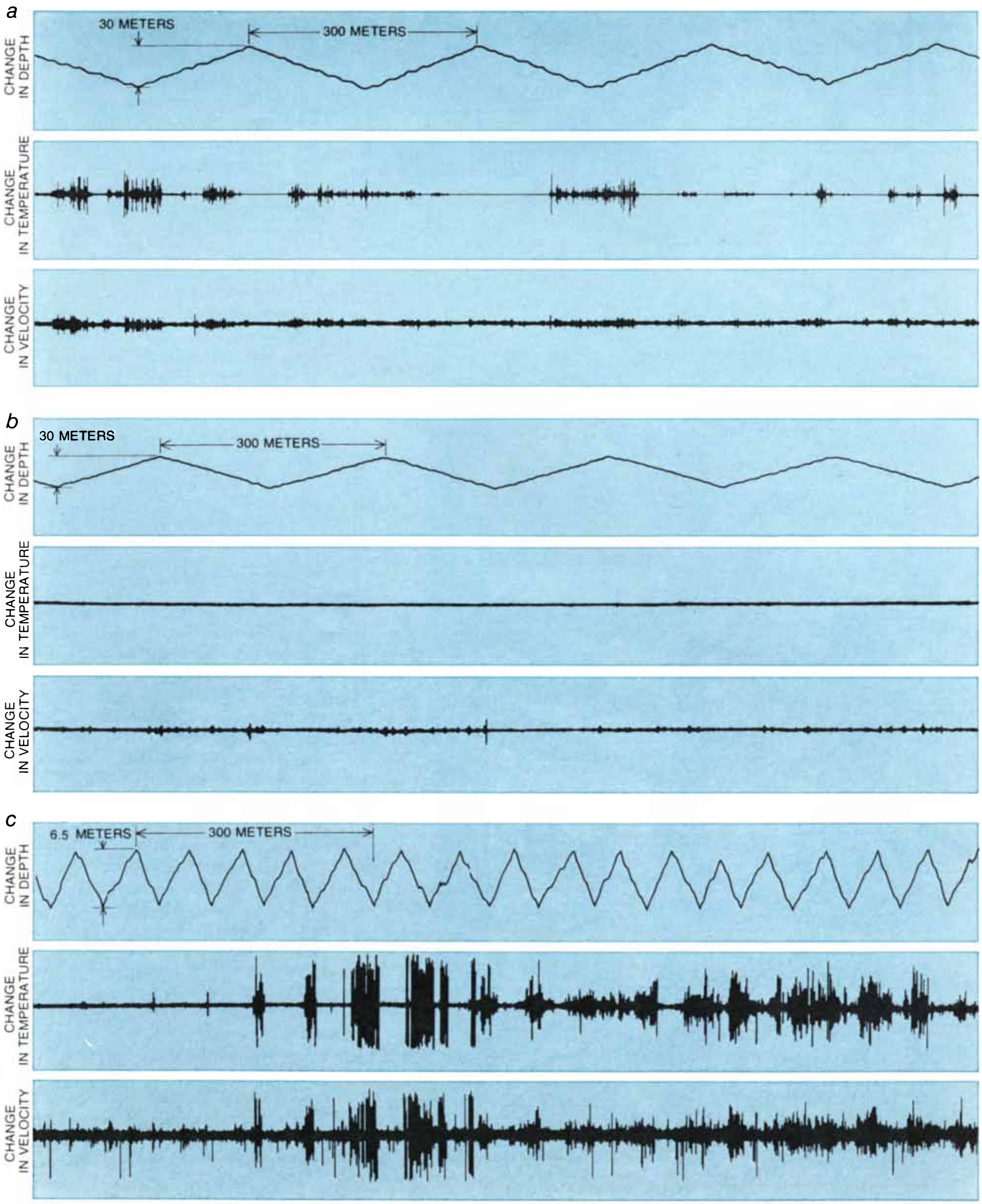


ALTERNATIVE THERMOCLINE MODEL minimizes the effect of turbulent stirring and attributes oceanic heat transport primarily to the development of subtropical gyres: slowly rotating mounds of water hundreds of kilometers across. Gyres are created by the prevailing westerly and easterly winds in conjunction with Coriolis force, which deflects moving bodies to the right in the Northern Hemisphere and to the left in the Southern. There would be a tendency for the warm surface water under a subtropical gyre to sink and flow outward.

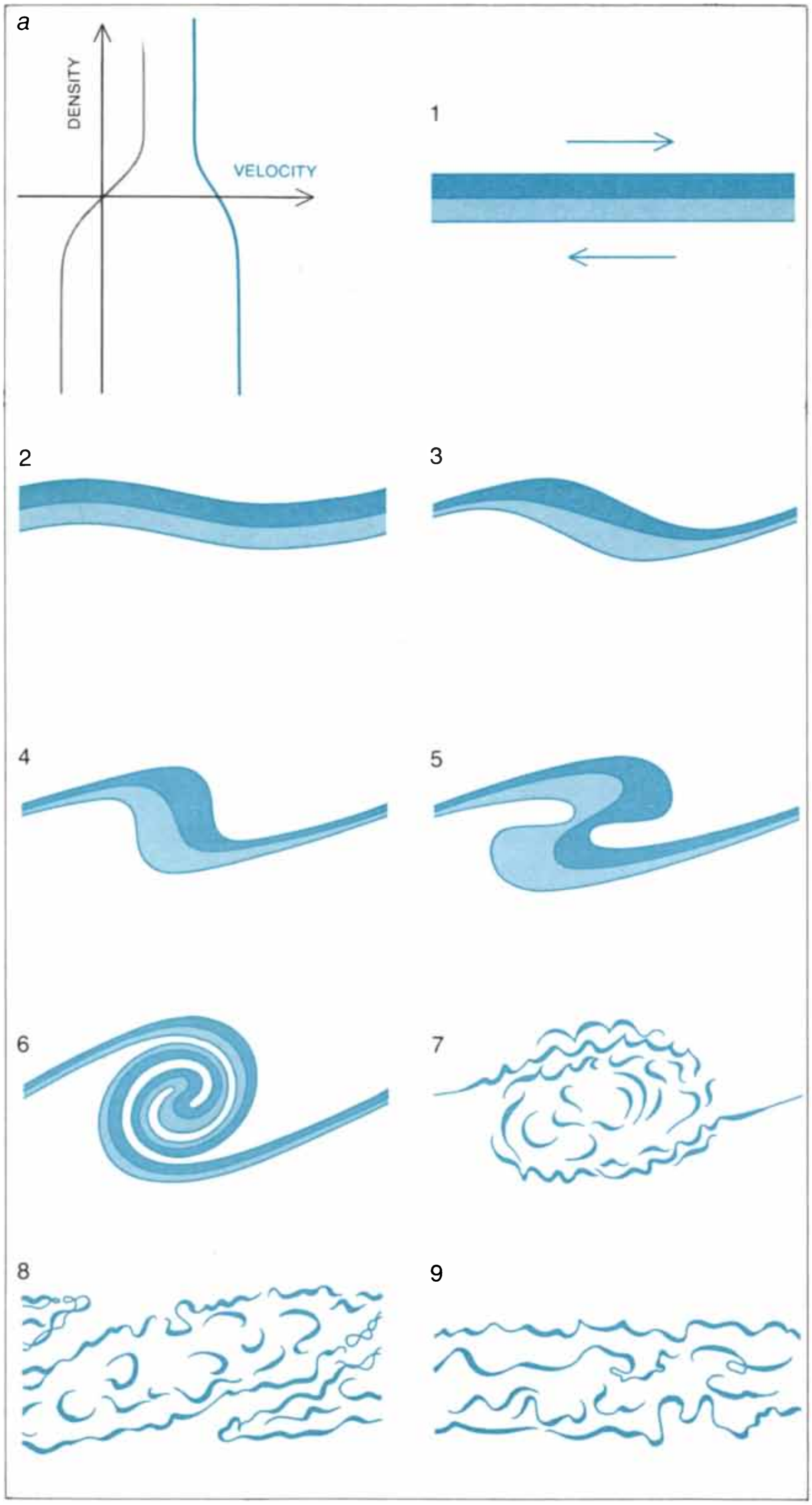
how the energy-dissipation rate and the viscosity can be combined in an equation that predicts the smallest-scale length for fluctuations of velocity. For the range of dissipation rates prevailing in the ocean the velocity cutoff scale—the dimension below which velocity fluctuations become insignificant—varies from about one centimeter to about six centimeters. Very little turbulent kinetic energy can exist over shorter dimensions.

If variations of salinity or temperature are present, they will be stirred by the same eddies that produce the velocity structure. The important result obtained from dimensional analysis is that the scale cutoffs for temperature and salinity are respectively only about a half and a fifth as large as the scale cutoff for velocity is. This means that in a sample of water in which velocity fluctuations had been smoothed nearly to zero, temperature fluctuations could still exist, and that if mixing continued until temperature fluctuations had been smoothed nearly to zero, salinity fluctuations could still exist. Although there is little kinetic energy associated with motions smaller than the velocity cutoff, the distortion of fluid elements by the remaining velocity shear will stir the temperature and salinity variations over smaller and smaller scales until ultimately they are both smoothed by diffusion. The importance for oceanography is that it is not necessary to resolve scales of variation smaller than the cutoff values in order to compute the rate of mixing in a turbulent flow field.

How does one estimate the energy-dissipation rates for the ocean on which the scale cutoffs are based? There are two principal forces acting on the ocean: the tides and the winds. Each has an average input of from five to 20 ergs per square centimeter of sea surface per second. It is believed that most of the tidal energy is dissipated in the shallow seas near the coasts, whereas most of the motions induced by the winds are dissipated in the upper layers of the open ocean. A typical value for the wind-driven energy input can be obtained by assuming that an average input of 10 ergs per square centimeter per second is dissipated uniformly in a mixed layer 50 meters deep lying above the thermocline. That works out to a dissipation rate of 2×10^{-3} erg per gram-second. A reasonable limit for the dissipation rate for the ocean below the thermocline can be estimated by combining the total tidal and wind energy and dividing it by the total volume of the sea. This gives a dissipation rate of 1×10^{-5} erg per gram-second. One can then compute the scale cut-



RECORDS OF HORIZONTAL MICROSTRUCTURE were made at three depths in the open Pacific off the west coast of Vancouver Island by Patrick W. Nasmyth as part of his doctoral thesis at the University of British Columbia. Changes in temperature and water velocity were sensed by probes on a submerged vehicle towed behind a ship. As the depth curves show, the vehicle was cycled up and down in porpoising fashion. The first series of curves (*a*) was made in a part of the ocean containing patches of turbulence, as indicated by rapid oscillations of the pens recording changes in temperature and velocity. The two traces generally coincide. The second series of curves (*b*) displays little microstructure activity, indicating that the ocean was quite stable at the recording depth. The last series (*c*) reveals the existence of a sharply defined front, or boundary, between two water layers; the boundary slopes upward about 1½ degrees with respect to the direction of the recording. At the outset the porpoising vehicle remains almost wholly within the upper layer, which is isothermal. Soon, however, it begins dipping into a strongly turbulent region. Owing to the slope of the front, the vehicle spends more and more time in the turbulent layer, until it is almost wholly immersed in it at end of trace.



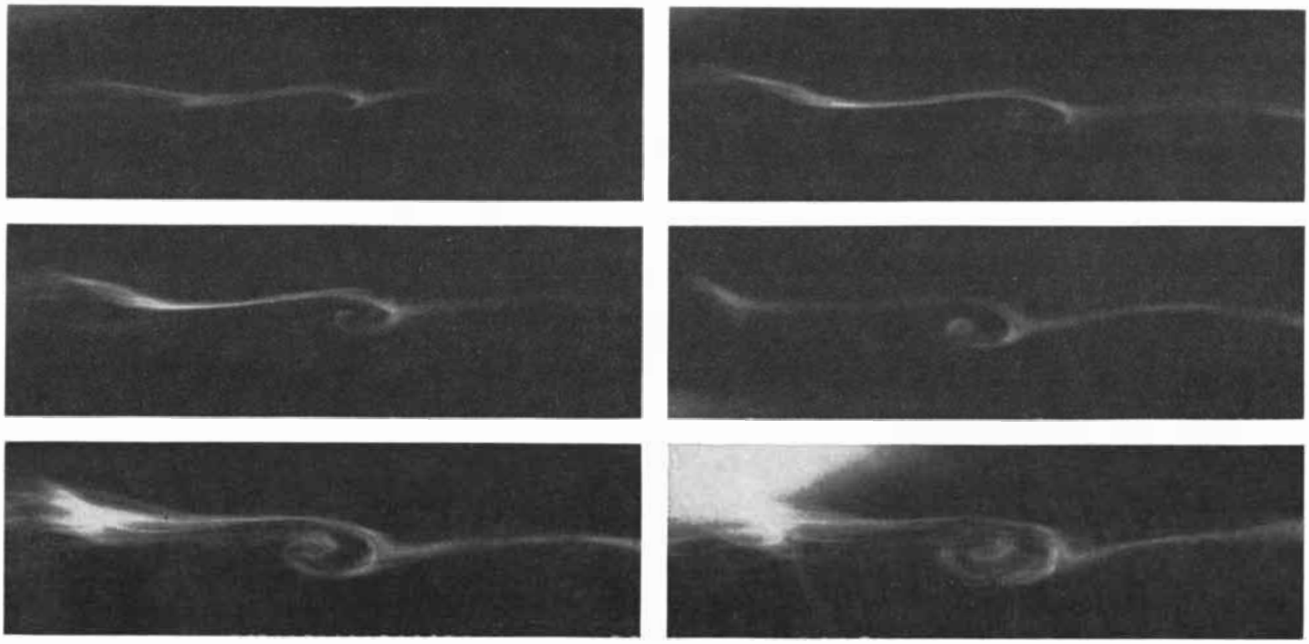
KELVIN-HELMHOLTZ SHEAR INSTABILITY is one form of process believed to stir the ocean in regions where high-frequency internal waves propagate between two layers of significantly different density. The formal situation is shown in *a*, where the horizontal arrow represents the interface between the two layers. The upper layer has a higher velocity and lower density than the lower layer. The following drawings are based on observations of the instability that can be generated in a two-layer laboratory tank by rapidly raising and then lowering one end of the tank. The two layers soon lose their coherence and break up into turbulent patches. The patches are rapidly flattened by stratification, which gives rise to a finely layered microstructure as the sloshing inside the tank subsides. The diagrams are based on a study by S. A. Thorpe of National Institute of Oceanography in England.

off for velocity, temperature and salinity based on these energy-dissipation limits. These dimensions establish the resolution the oceanographer must try to achieve in his measurements. Although a velocity probe must be capable of sensing fluctuations over several centimeters and a temperature probe must be capable of sensing them over about one centimeter, a salinity probe should have a resolution measured in millimeters.

Microstructure measurements capable of resolving the length-scale cutoff for both velocity and temperature have been made in the upper 300 meters of the ocean by a group of Canadian oceanographers, Harold L. Grant, Robert W. Stewart and Anthony Moilliet, who mounted sensitive probes on a vehicle towed horizontally behind a ship. This work has been continued by Patrick W. Nasmyth in the open coastal waters west of Vancouver Island. He has often found the upper 30 meters of the Pacific to be completely turbulent, with energy-dissipation rates as high as 2.5×10^{-2} erg per gram-second. The probes used are sensitive enough to distinguish between the cutoff length for velocity and the shorter cutoff length for temperature in water that is sufficiently turbulent. A probe capable of resolving the still finer cutoff length for salinity has yet to be operated in the ocean.

The source of the nearly uniform turbulent activity documented by the Canadian workers is the wind blowing on the surface of the sea. During strong storms the wind is capable of churning the surface layers to depths of tens of meters. If the intense stirring persists for a few days, it can drive the mixing nearly to completion, producing a highly uniform region, termed the mixed layer, extending 10 to 100 meters below the surface. The weak turbulent motions generated by only moderate winds are then sufficient to maintain the mixed layer for a considerable period.

Below the mixed layer the effects of surface motions are strongly attenuated. Intense turbulence arises only sporadically when there happens to be a local concentration of energy great enough to overcome the normal stratification, creating patches of microstructure activity separated by quieter water [*see illustration on preceding page*]. Dissipation values of 5×10^{-4} erg per gram-second have been recorded in patches at a depth of 90 meters. In addition to the regions of active turbulence one sometimes finds patches of "fossil" turbulence in which the velocity fluctuations have been dissipated by viscosity but the temperature



INTERNAL WAVES and Kelvin-Helmholtz shear instabilities were made visible by use of a fluorescein dye in the shallow waters off the island of Malta in the Mediterranean. These photographs, like the one on page 64, were made by Woods and a team of Royal Navy divers. The waves shown in the photographs have a length of 250 centimeters and a crest-to-trough height of 60 centimeters.

gradients have not yet been smoothed. High-resolution measurements of salinity would presumably show even older turbulence fossils in which the temperature as well as the velocity fluctuations have largely vanished.

Vertical observations of temperature and salinity that fully resolve features down to a centimeter or so and respond to fluctuations of millimeter size were first made with a freely sinking probe built by Charles S. Cox in our laboratory at the Scripps Institution of Oceanography. Temperature records obtained in mid-ocean at a depth of 1.5 kilometers appear almost as irregular as those from lesser depths, although both the amplitude and the sharpness of the fluctuations are greatly reduced. One of the common patterns shows intense turbulent fluctuations similar to the patches disclosed by the horizontal measurements. Even though we have not measured velocity fluctuations we can infer from the sharpness of some of the individual temperature gradients, which are less than a centimeter thick, either that they were being maintained by continuous stirring or that they had been formed only a few minutes before the observation. Molecular diffusion will smooth an infinitely sharp temperature step to a thickness of one centimeter in three minutes.

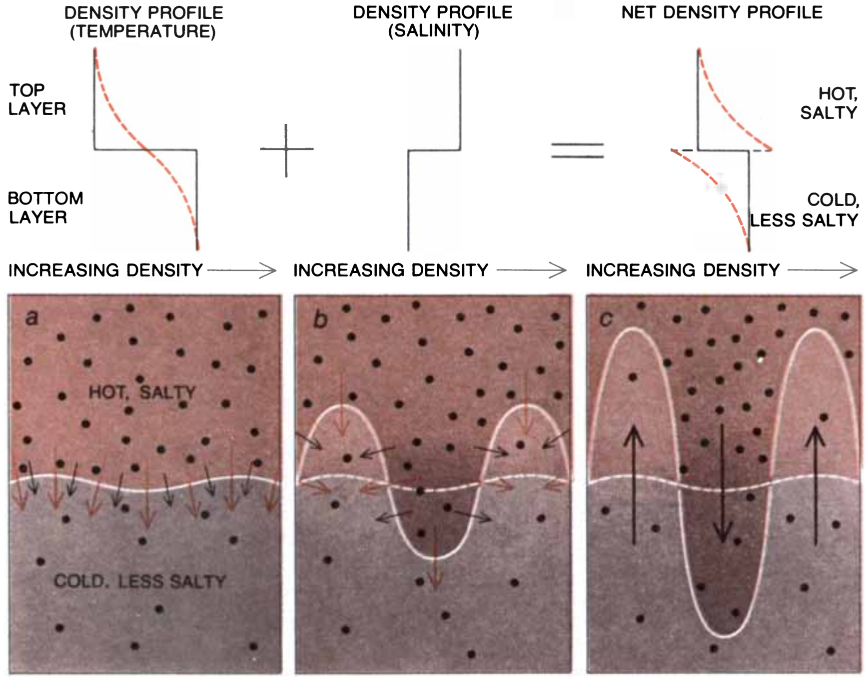
As part of my own graduate research I built a probe that senses electrical conductivity. This device has been used in concert with the more recent observations, making it possible to calculate the salinity and hence the density on a centimeter scale. We have observed local density inversions that cannot exist as long-term features and thus must represent transient overturns of the water. The concentrations of energy capable of producing such turbulent patches often indicate the presence of internal, or subsurface, waves.

Internal waves propagate energy through the body of a stratified fluid in much the same way that energy is propagated by waves at the surface. The ability of the sea surface to sustain undulating motions is a consequence of the buoyant restoring force that results from the density differences between air and water. Density contrasts below the surface are considerably weaker and occur as continuous changes across interfaces that are much more diffuse than the air-water boundary. As a result internal waves usually have greater amplitudes and longer periods than those on the surface. The sharpest density transition in the ocean is at the bottom of the mixed layer; a more diffuse transition coincides with the region of rapid temperature decrease in the thermocline. Internal waves with an amplitude of about 10 meters commonly travel through the thermocline with periods of about 20 minutes.

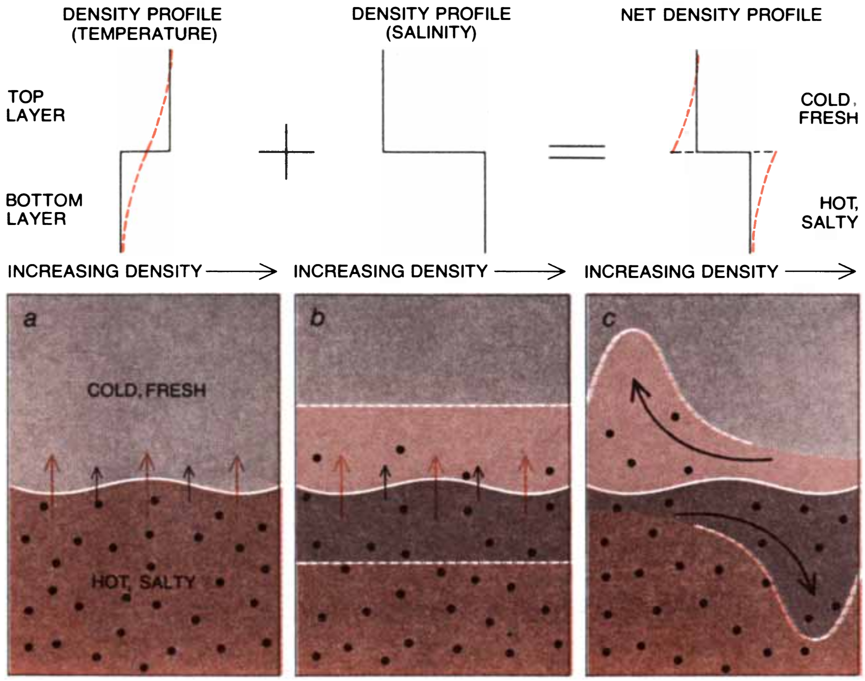
At least some degree of internal wave motion appears to exist throughout the ocean. Although the energy for these waves is ultimately derived from the tides and the winds, the manner in which the energy is converted into internal oscillations is poorly understood. Two principal mechanisms are thought to be involved. The first is the flow of steady currents and tidal currents over irregular contours on the sea floor, which would induce local oscillations in the mass of water flowing near the bottom. The second likely mechanism is the resonant interaction of surface waves, which transfers some energy to internal waves. Given the weak stratification present in the deep ocean, internal waves can propagate on sloping paths, providing a way to radiate energy from the region of generation into the interior of the ocean.

Internal waves do not result in significant energy dissipation unless they lead to an instability. One form of instability, known as Kelvin-Helmholtz instability, tends to arise when high-frequency waves propagate along rather sharp density discontinuities. Growing slowly at first as small wavelets on the crests of an internal wave, the instabilities rapidly form rolls that show much microstructure activity. Laboratory studies conducted by S. A. Thorpe at the National Institute of Oceanography in England have shown that the structures finally break down into turbulent patches that are rapidly flattened by the surrounding stratified fluid, producing a finely layered region as the stirring motions subside [*see illustration on opposite page*].

Using dye tracers in the shallow water



SALT-FINGERING is a special type of vertical-transport process that can arise because heat diffuses about 100 times more rapidly than salt. Salt-fingering occurs in a two-layer system where the upper layer is warmer and saltier than the lower layer, as is shown in the density profiles at the top. Solid lines show the initial density profile, whereas the broken lines show the situation a few minutes later when the more rapid diffusion of heat has produced a potentially unstable situation at the interface. The three lower panels show how small irregularities on the interface lead to salt-fingering. Color represents temperature and heat flow; black dots and arrows indicate salt and salt diffusivity; gray tones indicate density. A downward bulge will cool the water above the interface more rapidly than when the interface is level or curves upward. The rapid loss of heat makes the water above the bulge heavier than its surroundings, so that it tends to sink. Similarly, the water below an upward bulge is warmed more than its surroundings and tends to rise. The result is fingering (*c*).



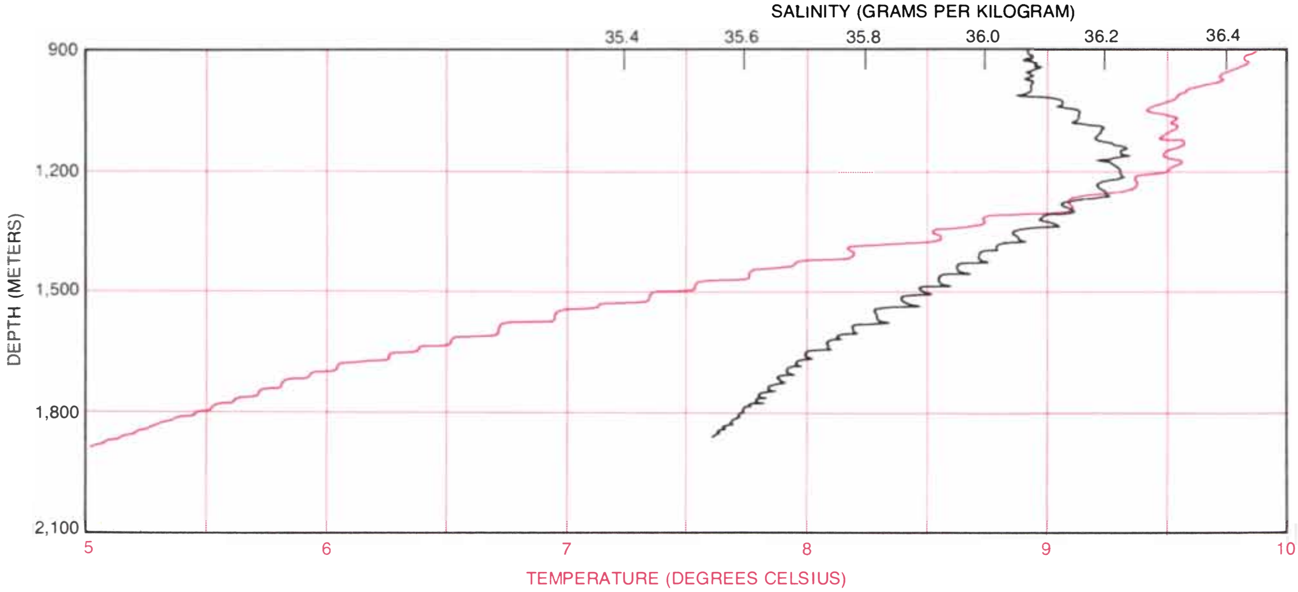
ENHANCED STABILITY of the density step in a two-layer system results when the upper layer of water is cool and fresh and the lower layer is warm and salty. Now as heat diffuses across the interface between the layers more rapidly than salt (*a*, *b*) the water that is immediately above the interface becomes even lighter than the upper layer as a whole and tends to rise, whereas water just below interface increases in density and tends to sink (*c*).

off the island of Malta, a team of Royal Navy divers under the direction of John D. Woods of the University of Southampton were able to observe and photograph internal waves and the occasional shear instabilities they generate [*see illustration on preceding page*]. The instabilities strongly resemble Thorpe's laboratory observations. Although the growth of the instability took only one or two minutes, the turbulence persisted for five or 10 minutes after breaking, and a dye scar remained for many hours. Temperature records that we have made at greater depths in the Pacific show distinctive S-shaped features over vertical distances of half a meter to three meters that suggest similar instabilities in their initial stages.

The mixing effectiveness of an individual instability is hard to determine but is probably rather low because of the rapid damping that follows. Data taken in coastal waters at depths down to several hundred meters suggest that instabilities due to internal waves occur at a given point between once and six times a day, and that their averaged effect is equivalent to a steady turbulent mixing coefficient of between half a square centimeter and one square centimeter per second, or roughly the amount required for mixing in the diffusive thermocline models. By collecting an extensive series of observations from mid-ocean locations it should be possible to estimate ocean-wide mixing rates as one way of testing the various thermocline models.

The other vertical transport process that we know of originates in molecular motions rather than in the meter-scale velocity shears produced by internal waves. The molecular motions arise because heat diffuses much more rapidly from point to point than salt does, with the result that adjacent layers of water can rapidly destabilize under certain conditions. In a two-layer system involving both heat and salt one expects the system to be stable whenever the upper layer is less dense than the lower layer. Depending, however, on the actual distribution of heat and salt, three distinct situations can arise: one inherently stable, one inherently unstable and one where the layers remain separate but where there is convective stirring in each layer.

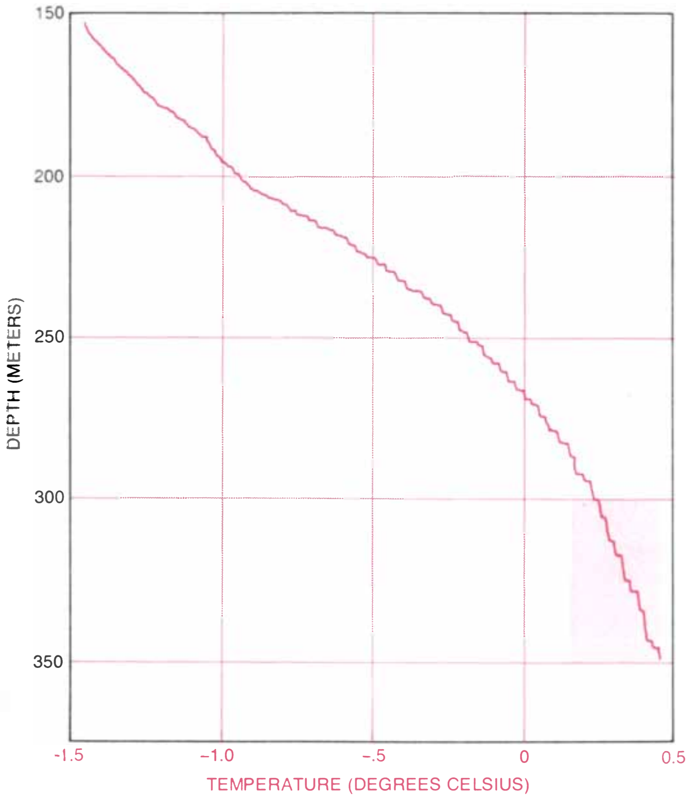
Laboratory experiments conducted by J. S. Turner of the University of Cambridge and Henry Stommel of the Woods Hole Oceanographic Institution showed that the inherently stable system is produced if the upper layer is both warmer

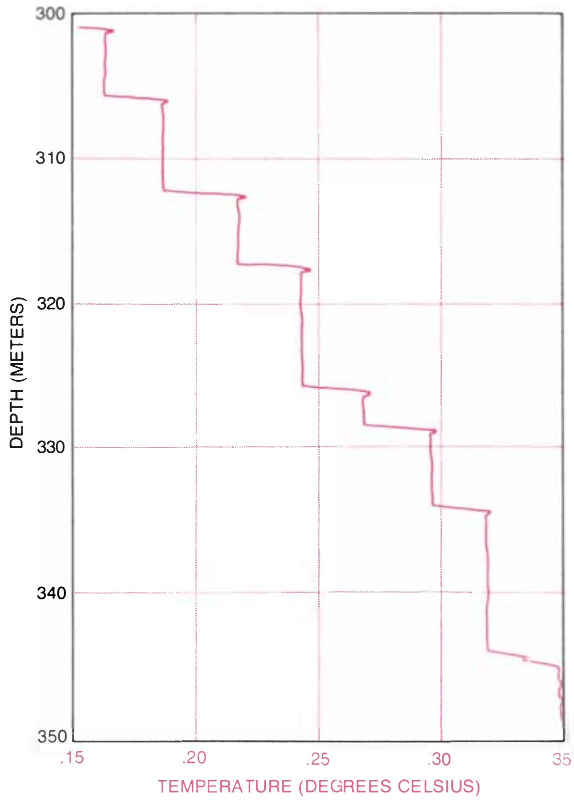


PROBABLE EXAMPLE OF SALT-FINGERING is represented by the series of temperature steps (*color*) and salinity steps (*black*) recorded in the Atlantic Ocean below an intrusion of warm saline water that has flowed westward through the Strait of Gibraltar. Horizontal layers may extend for 30 miles. Recordings were made by R. I. Tait and M. R. Howe of the University of Liverpool.

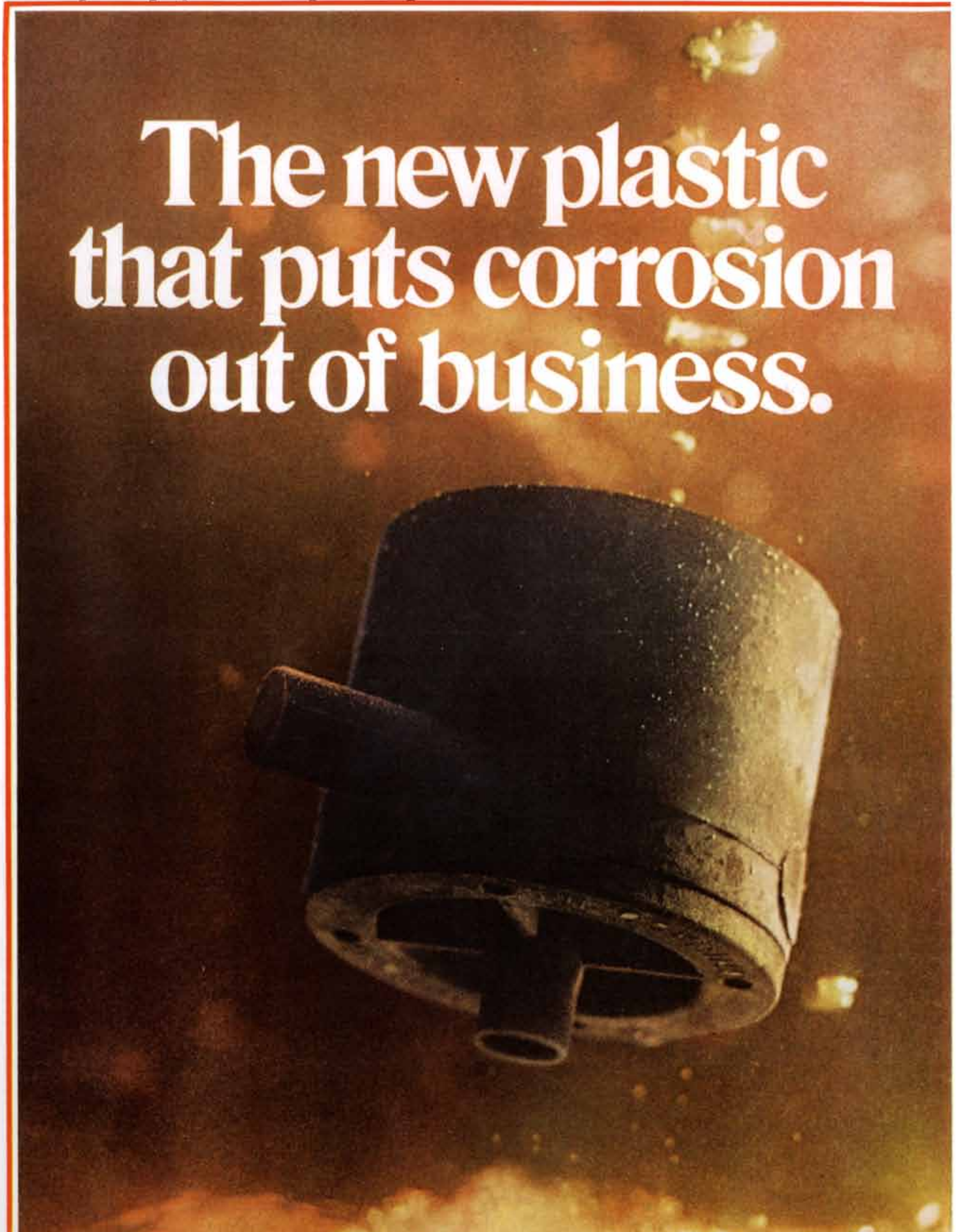
and less salty than the lower layer; both components act together to produce a stable density step across the interface. The inherently unstable system is produced if the upper layer is warmer and saltier than the lower layer, provided that the lower layer is heavier primarily because of its cooler temperature [*see top illustration on opposite page*]. The third system, which leads to convective stirring in each layer, is produced if the upper layer is cooler and less salty than the lower layer, provided that in this case the lower layer is heavier primarily because of its salinity [*see bottom illustration on opposite page*].

If the last two systems are set up experimentally with sharp interfaces, the temperature step will be smoothed by diffusion in just a few minutes, whereas the salinity step will be virtually unal-





SERIES OF TEMPERATURE STEPS was observed in the Arctic Ocean under the floating ice island designated T-3 by Steve Neshyba and Victor T. Neal of Oregon State University and Warren Denner of the U.S. Naval Postgraduate School at Monterey, Calif. Although the corresponding salinity steps were not recorded, it seems likely that the temperature steps are another example of stratification produced by the unequal diffusivity rates of heat and salt. The layers were formed above an intrusion of warm saline water from the Atlantic. The region in rectangle at left is enlarged in the curve at the right. Seven steps are compressed within .2 degree C.


The new plastic
that puts corrosion
out of business.

tered. The reason is that heat diffuses about 100 times more rapidly than salt. In the first case, where the warm, salty water is on top, the rapid loss of heat across the interface makes the water immediately above the interface heavier than the less salty water immediately below it, which is simultaneously absorbing heat and becoming lighter. The result is a convection pattern resembling a tiny checkerboard, in which alternate cells of salty water sink while adjacent cells of less salty water rise; the pattern is known as salt-fingering. In the second case, where warm, salty water is on the bottom, the rapid diffusion of heat upward increases the stability across the interface but produces strong convective stirring within each of the layers.

The salt-fingering pattern of convection persists only for a limited depth below the interface. At a certain point the negative-buoyancy flux due to the descending saline water forms a thick region that undergoes a weak, overturning turbulent motion, thus creating a well-mixed layer whose temperature and salinity are intermediate between those of the two original layers. In a deep column of water a series of thin salt-fingering regions and thick well-mixed layers can form progressively below the original interface, giving rise to stairstep profiles of decreasing temperature and salinity. In a similar manner the convective motions above a layer of warm saline water are self-limiting. In this case a family of steps is progressively formed above the original interface. A vertical profile through such a system shows steps of increasing temperature and salinity with increasing depth. It appears that the same sequence of steps can develop in the absence of a sharp interface if the temperature and salinity gradients are in the proper sense, evidently as a consequence of small irregularities in velocity that distort the original gradients to produce regions of increased heat diffusion.

In the ocean there are many areas over which the temperature and salinity profiles are in the right relation for differential diffusion to occur. As many as 20 steps of decreasing temperature and salinity, apparently the result of salt-fingering, have been observed in the Atlantic under an intrusion of Mediterranean water that moved through the Strait of Gibraltar [*see top illustration on page* 75]. An even greater number of steps of increasing temperature and salinity, each about three meters thick, have been discovered in the Arctic Ocean above a layer of warm saline water that has entered from the Atlantic [*see bottom illustration on page* 75]. This is an example of the second case, in which steps are progressively formed above an intrusion of warm saline water. On a smaller scale the upper sections of temperature inversions 10 meters thick found off San Diego have been resolved into families of layers 20 to 30 centimeters deep separated by descending steps of increasing temperature and salinity that are often less than two centimeters thick.

By producing convective motions, which drive turbulent stirring within layers, and by creating steep gradients of temperature and salinity at layer boundaries, differential diffusion produces vertical transport rates that are much greater than those that would result from simple molecular diffusion along the average gradients. Differential diffusion is clearly a major mixing process in places where warm saline water mingles with cooler fresher water to produce large-scale profiles of high variation. The upward fluxes of temperature and salinity computed for the thin diffusion layers off San Diego are great enough to fill in the minimums of temperature and salinity between successive intrusions in about a day. This suggests that the boundaries of the intrusions are rapidly altered until they are no longer recognizable as separate layers. It remains to be demonstrated whether or not salt-fingering is a general process in the deep ocean. An effort to use optical means for observing centimeter-diameter salt fingers in the open ocean is being made by Albert Williams III of Woods Hole. If salt-fingering is present, it will complement the instabilities created by internal waves and other shearing processes in supplying the downward transport of heat required to maintain the thermocline profile.

Although I have discussed differential diffusion and shear instabilities as separate processes, in the ocean they must interact and modify each other in ways that now can only be guessed at. For example, laboratory work has shown that salt-fingering is altered and partially inhibited by the presence of turbulence or even by weak shearing currents. Conversely, the sharp interfaces that are formed by both types of differential diffusion process must inevitably modify the character of internal waves. By advancing our understanding of the small-scale dynamics of the ocean, observations of oceanic microstructure can provide tests of the validity of the various thermocline models and ultimately improve our ability to predict the dispersal of substances introduced into the sea by both nature and man.

The Crashworthiness of Automobiles

The statistics of automobile crashes suggest changes in automobile design for the prevention of death and injury. Among these changes are a crushable front end and a reinforced passenger compartment

by Patrick M. Miller

Can the automobile be made significantly safer? Given the current highway conditions, the variety of public tastes in motorcars and the wide range in drivers' skills and habits, can a car be designed that would reduce the high toll of deaths and injuries on the highways by providing more protection to drivers and passengers in collisions? Studies in the program of "basic research in crashworthiness" sponsored by the National Highway Traffic Safety Administration (NHTSA) indicate that several structural ideas now under test look promising. I shall discuss the present appraisal of the situation on the basis of tests conducted by the Cornell Aeronautical Laboratory, Inc. (now the Calspan Corporation), which for many years has been investigating automobile accidents and safety measures.

Although automobile safety has been a concern almost since the invention of motorcars, it was not officially recognized as a national problem until 1966, when the traffic and motor vehicle safety act was enacted by Congress as a consequence of an alarming increase in the number of highway deaths. The death rate per 100 million miles driven had leveled off, but as Americans took to the road more and more, the total numbers of accidents, deaths and injuries rose sharply. The National Safety Council estimates that the yearly totals are now 16 million accidents, 55,000 deaths and about two million at least temporarily disabling injuries.

The NHTSA has instituted several programs, including some directed to drivers and the care of accident victims; the major effort, however, is focused on the automobile itself, with new motor vehicle safety standards imposed on the car manufacturers. The problem turns out to be far more complicated than one might expect.

In designing a safety system one must start by defining precisely what it is intended to guard against. The NHTSA laid down as the basic specification that the crashworthiness of a vehicle is to be measured by its ability to protect riders and pedestrians from injury, not in terms of the cost in damage to the vehicle itself. This, however, is only the first of many questions that need to be decided in arriving at an appropriate design.

For example, one must consider the protection not only of the riders in a vehicle but also of the occupants of another car with which it may collide. A strong, heavy front end with high capacity for absorbing energy might result in an aggressive vehicle that was a hazard to any other car while protecting its own occupants. Furthermore, in the interest of saving lives a crashworthy vehicle must be generally compatible with the great range in size, weight and performance of the 110 million vehicles now on the nation's streets and highways.

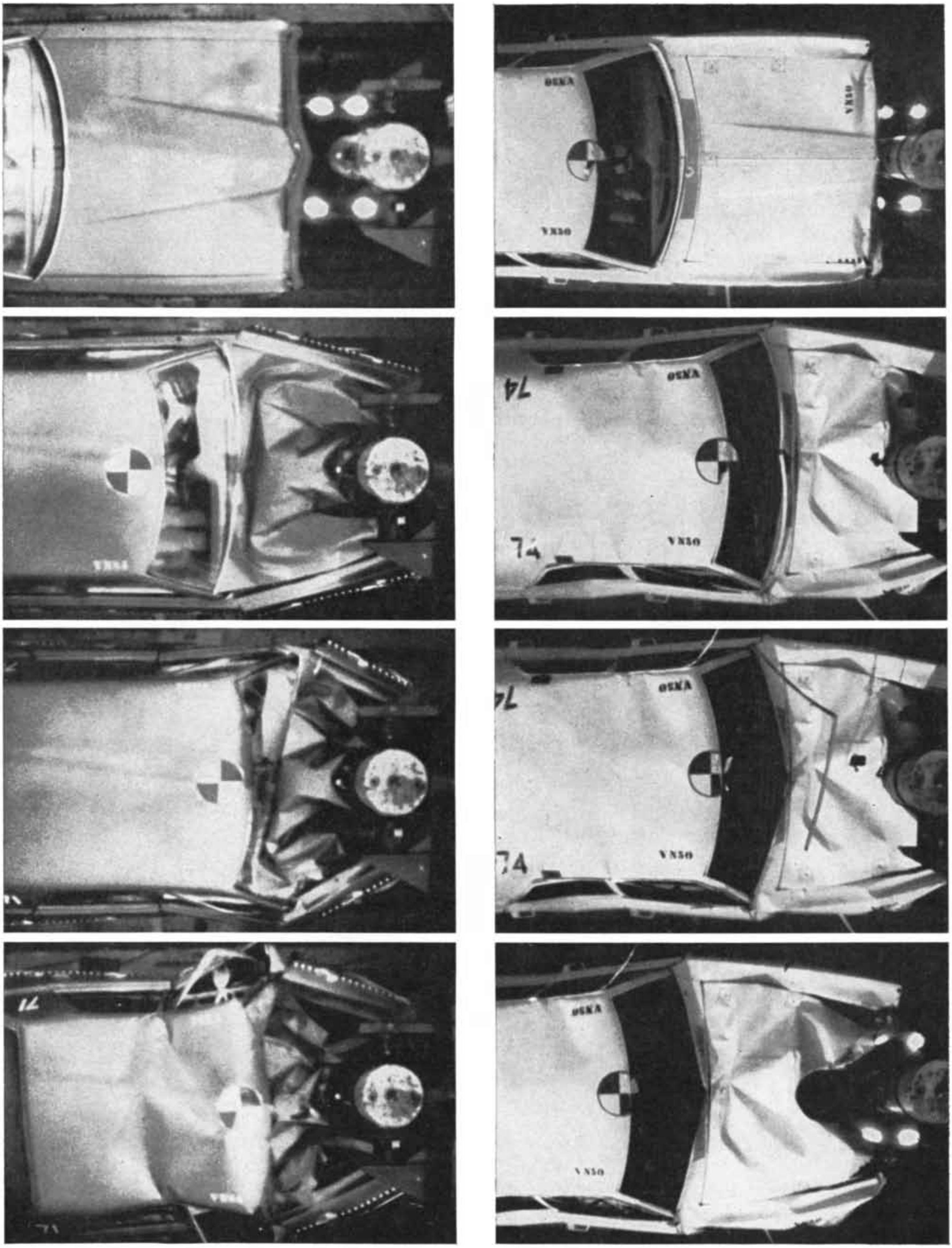
Another complicating factor to be considered is the distinct difference between the accident pattern in rural areas and that in urban areas. For example, more than two-thirds of the fatal accidents occur in rural areas (where the mean impact speed of the automobile in such accidents is estimated to be approximately 50 miles per hour); on the other hand, the urban areas, with their higher population density, account for nearly two-thirds of the nonfatal injuries from automobile accidents. Roll-over accidents are much more common in the country than in cities; conversely, accidents in which one vehicle strikes the side of another are more common in the city, with its many intersecting streets.

Still another consideration is the need to give better protection to pedestrians. In 1971 about a fifth (10,600) of the people killed by automobiles were pedestrians, and approximately 150,000 pedestrians were injured. The automobile hazard for pedestrians was of course much more acute in urban areas than in rural ones.

Designing a vehicle that would leave a pedestrian unscathed after hitting him is obviously an all but unsolvable problem. Some improvements that would tend to reduce the severity of the impact are possible, for example eliminating protrusions that might snag the pedestrian and relaxing the stiffness of the hood and grille structures. On being hit by a vehicle, however, a pedestrian often suffers two impacts: one when the vehicle strikes him and another when he strikes the road surface. There does not seem to be any way to avoid or soften the second impact. On the whole, it appears that efforts for the protection of pedestrians must be directed mainly to the design of traffic-control systems that will separate automobiles and pedestrians to the greatest possible extent.

To deal with the contrasting conditions in urban and rural driving, it seems reasonable to consider developing specialized designs for crashworthiness, one for automobiles that will be used predominantly in the urban environment, the other for rural use. Investigation of this possibility may well prove to be rewarding in reducing the accident toll.

For the present, however, investigators and designers are concentrating on problems and projects that must be given the highest priorities. In a field that is so complicated and has been so little explored, it is the better part of wisdom to begin by establishing a rational basis of understanding and developing the state of the art step by step. For understanding the problem and testing possible answers the most useful information we have available is the accumulation



TEST CRASHES of a conventional automobile (*left*) and a structurally modified automobile (*right*) were carried out on the grounds of the Cornell Aeronautical Laboratory (which is now the Calspan Corporation). The photographs were made with a motion-picture camera mounted above the crash site as the vehicles were towed along a track so as to crash into a heavy steel pillar at a speed of 58 miles per hour. The modified automobile, which had a reinforced frame and front end and an arrangement that prevented the motor from being driven into the passenger compartment, suffered appreciably less damage in the passenger compartment.

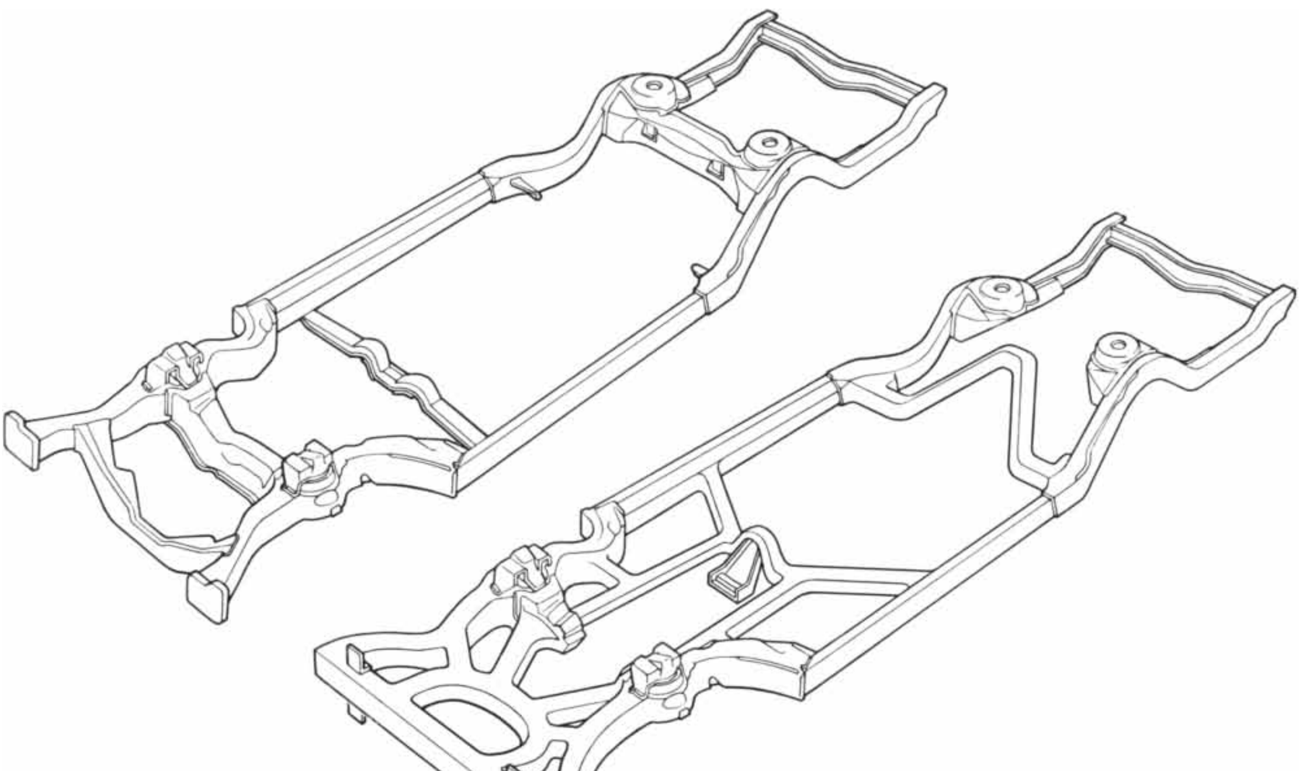
of statistics on automobile accidents. Searching for patterns in these data, we can see suggestions for specific devices to make automobiles more crashworthy.

Certain significant patterns were found in an analysis of approximately 25,000 injury-causing accidents, predominantly in rural areas, recorded in the files of the Calspan Corporation's research project on injuries in automobile crashes. The accidents were classified according to the angle of collision (frontal, rear or side), and there was a separate category of accidents in which the vehicle rolled over with or without collision. The most interesting finding from the analysis of these statistics was a showing that although frontal collisions are the most frequent accidents, side collisions and roll-overs are by far the most dangerous, as measured both by the extent of hazardous damage to the automobile and the actual occurrence of severe or fatal injury to the driver. Whereas side impacts accounted for only 13 percent of all the accidents, 23 percent of them resulted in extensive car damage and 16 percent resulted in dangerous or fatal driver injuries. Similarly, of the roll-overs (approximately 21 percent of all the accidents) 23 percent entailed heavy damage and 17 percent caused dangerous or fatal injuries to drivers. On the other hand, although frontal collisions were the most common type of accident (59 percent), only 9.5 percent of them entailed heavy damage and 9.6 percent resulted in death or dangerous injury to the driver [*see illustration on page 86*]. Only a small proportion of the 25,000 injury-producing accidents (4.5 percent) were of the type in which the vehicle is hit in the rear, which suggests that protection for the rear of the car will not require as much attention as protection for the front and sides.

Overall, 11.4 percent of the 25,000 injury-producing accidents resulted in dangerous or fatal injury to the driver. About half of these serious incidents were frontal collisions and the other half side collisions and roll-overs. Consequently in designing for crashworthiness, protection in side or roll-over accidents will require as much attention as is given to protection in frontal collisions. In the side-collision and roll-over cases there is a particularly high risk of being thrown from the car, which means emphasis must be placed on lateral restraints and measures to prevent the doors from becoming portals for ejection.

In order to evaluate the effectiveness of devices for protection of the automobile occupants in collisions it has been necessary to estimate the limits of the human body's tolerance for deceleration. Experiments with dummies cannot give a real measure of human responses. It can safely be assumed, however, that the severity of the exposure will depend on the level and duration of the deceleration. On this basis the Society of Automotive Engineers has formulated a "severity index" for exposure of the head, and the limited information from accidents and tests suggests that an exposure greater than 1,000 on this index may be extremely dangerous to life.

The NHTSA has set forth certain safety standards. It specifies that in experimental vehicles the deceleration of the head must not exceed 80 *g*'s (80 times the acceleration of gravity) if it is continued for more than three milliseconds. For future production vehicles the NHTSA standards limit a modified head-severity index (termed head-injury criteria) to 1,000, the deceleration of the



FRAME MODIFICATIONS made by the Calspan group to improve the crashworthiness of an automobile are depicted in a comparison of an unmodified frame (*top*) and a modified one (*bottom*) for a standard-size car of U.S. manufacture. The modifications included removing some parts of the frame and adding other parts, all with the objective of increasing the energy-absorbing capability of the front and sides of the car and of providing for the engine to be deflected toward the ground in the event of a frontal collision.

chest to 60 g's for three milliseconds and the dynamic pressure loading on the upper femur (hipbone) to 1,400 pounds. These criteria are being tested on dummies ranging in size from extremely small girls to extremely large men.

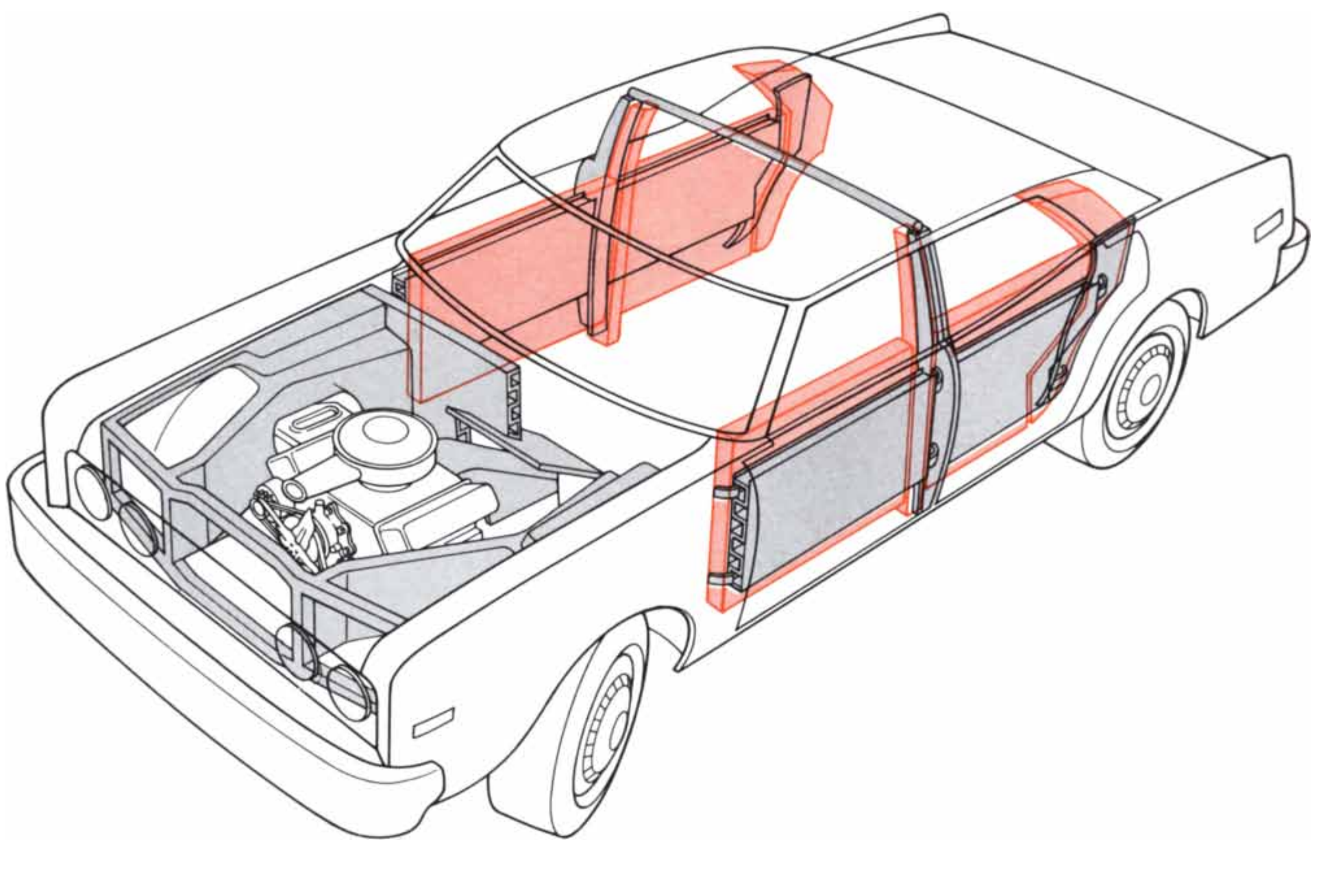
How does the structure of present automobiles perform in protecting occupants from dangerous levels of deceleration? This question, among others, has been studied in more than 100 automobile collision tests performed by our laboratory during the past four years. The results of the tests indicate that the structures of current full-size automobiles are capable of keeping the deceleration of an adequately restrained body in the passenger compartment within tolerable limits in a frontal collision with a fixed object up to an impact speed of between 35 and 40 miles per hour, in a side collision with a tree or utility pole at an impact speed of up to about 10 miles per hour and in a side collision with a vehicle of similar size at a closing speed of up to about 30 miles per hour. In such collisions (typical of accidents in the urban environment) there was no serious intrusion of one vehicle into the other, nor were the doors forced open.

Even at these moderate speeds, however, the dummies representing human occupants experienced what would plainly be injurious impacts with the interior structures in the automobile. Hence it is clear that the first priority in improving protection for the driver and passengers in current automobiles should be the provision of suitable systems to restrain the dislodging of the occupants from their seats and to buffer the effects of impacts with obstructions and glass in the vehicle.

In this connection it is instructive to study the experience with seat belts. Such an analysis was made by our laboratory on a sample of accidents in Utah involving 14,261 occupants, 2,359 of whom were wearing seat belts at the time of the collision. The analysis showed that failure to use the seat belt increased the risk of death or severe injury by 100 percent. The ratio of those who received nontrivial injuries was 70 percent higher among the unbelted than among the belted, and the risk of incurring at least some injury was 40 percent greater among the unbelted than among the belted. Three-fourths of the benefit of wearing the seat belt could be attributed to protection against injury within the vehicle, and the remaining one-fourth to a reduced probability of being thrown from the car. The study indicated that seat belts were most useful in the cases of side collisions.

In all likelihood the combination of the seat belt with the shoulder strap will be found to be even more effective. Yet the fact must be faced that most motorists do not make use of these restraint systems. In the Utah statistics (from 1969) only 16.5 percent were wearing the seat belts; the estimate at present is that, notwithstanding all the urgings by authorities, only about 25 to 35 percent use the lap belt and only about 5 percent the lap-and-shoulder combination. Consequently investigators working on the problem of automobile safety have turned to the other alternative: designing "passive" restraint systems that call for no effort or choice on the part of the motorist.

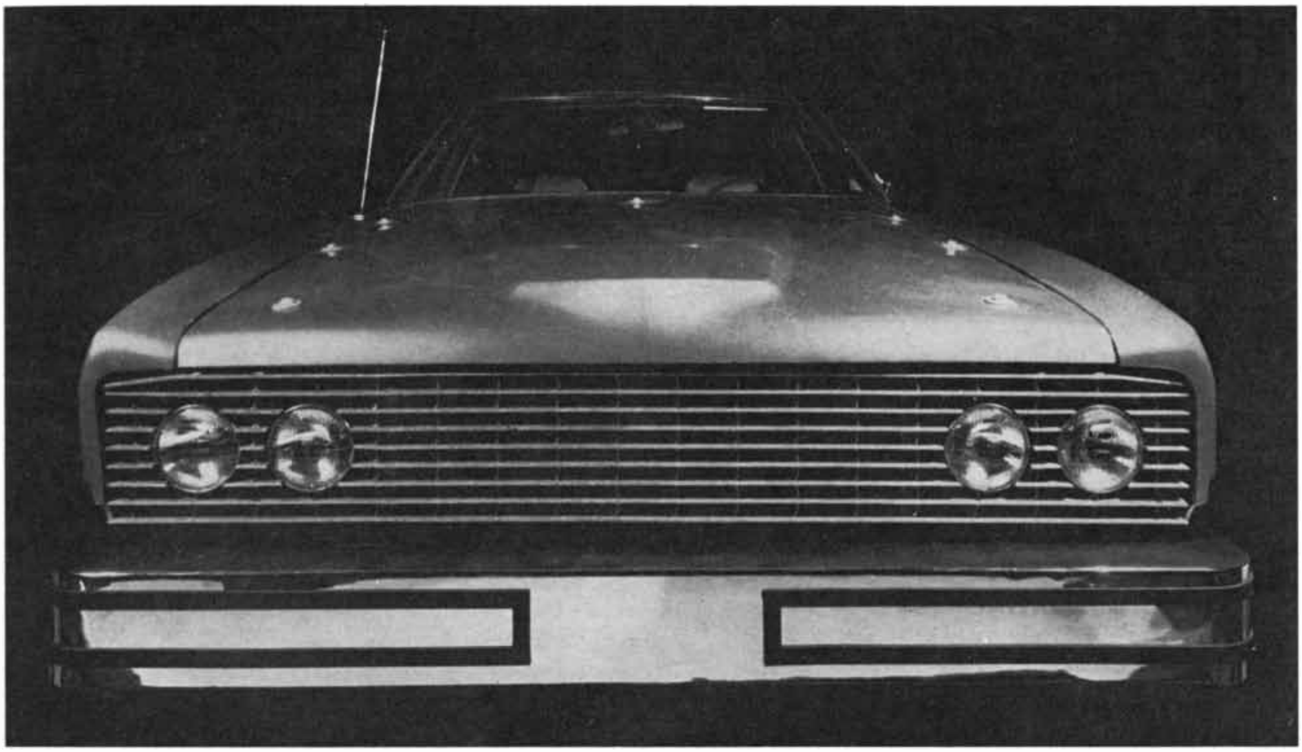
This line of research has focused mainly on the air bag or inflatable cushion that is automatically inflated in front of the car occupant as soon as the system senses a frontal collision. Estimates of



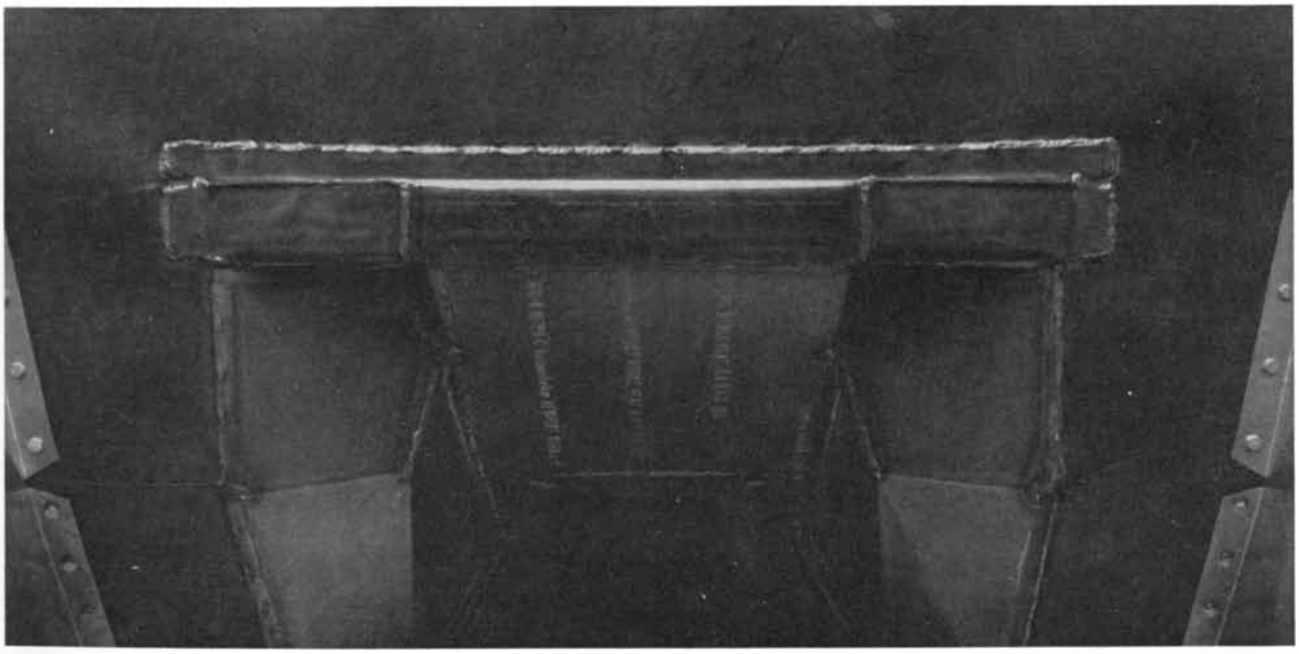
MODIFIED AUTOMOBILE built on the modified frame shown in the illustration on the opposite page includes changes in the front end, the engine mounting and the exterior side panels (*gray*) and the addition of several inches of energy-absorbing material to the interior panels of the passenger compartment (*color*). The slanting structure at the rear of the engine is part of the arrangement made to deflect the engine toward the ground, rather than into the passenger compartment, if the automobile undergoes a head-on collision.

the usefulness of such a system must take into account, among other things, the time available for the system's deployment in an accident. This time depends on two factors: (1) the occupant's distance from interior structures with which he may collide during the vehicle's collision and (2) the distance of the crushing in of the exterior structure. As this structure is being crushed, the occupants' compartment continues to move forward in the original direction of travel until the entire vehicle is stopped. Hence the forward displacement of the compartment during the collision, measured by the extent of crushing of the exterior structure, adds to the distance available for stopping an occupant's motion before impact with, for example, the instrument panel.

Assuming that the average decelera-



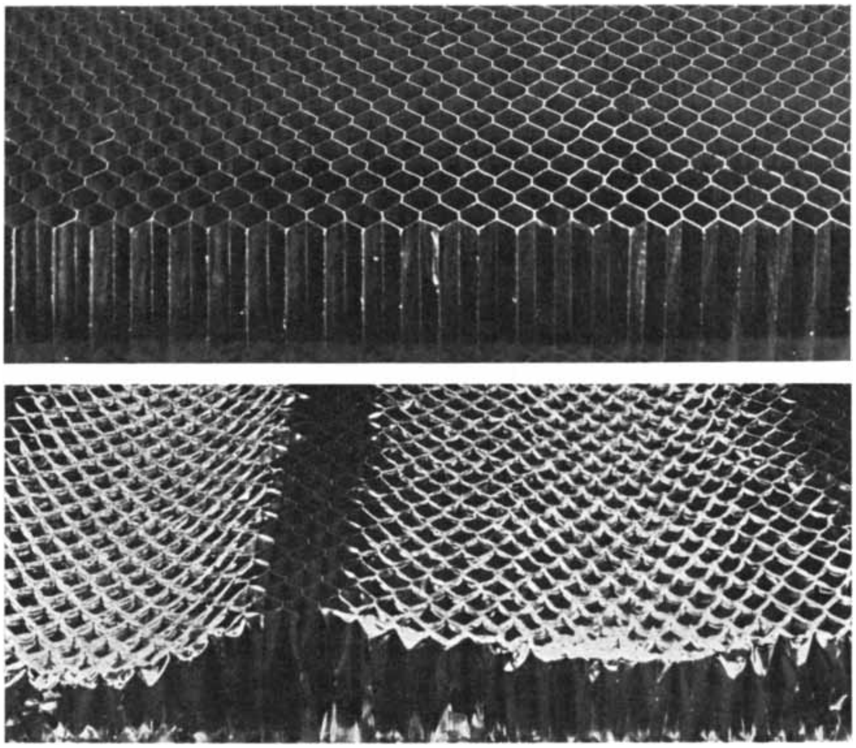
FRONT-END MODIFICATIONS designed by the Calspan group include a strong bumper made of high-strength steel and backed with foam urethane and an aluminum grille built to absorb more energy than the grille of a standard automobile does in a collision.



FIRE WALL normally provided between the engine and the passenger compartment is modified by the Calspan workers to provide a slanting channel that, in a crash, directs the engine downward and causes the passenger compartment to ride up over the engine.

tion of a car in a crash is 20 g's, it can be calculated that this displacement in a current full-size automobile colliding head on with a flat barrier at a speed of 40 miles per hour is 32 inches. Assume that the occupant is 20 inches behind the instrument panel. The total available distance for checking his movement then is 52 inches. In present air-bag systems it takes at best about 50 milliseconds to deploy the system: approximately 15 milliseconds for sensing the collision and 35 milliseconds to inflate the bag. During this time the occupant will have traveled forward 35 inches (since he continues moving as an essentially free body at the precollision velocity of 40 miles per hour). This leaves an effective distance of only 17 inches for stopping his motion before he reaches the instrument panel. In other terms, of the 91 milliseconds it takes for the vehicle to come to a full stop in a 40-mile-per-hour collision, only about 30 milliseconds is available for effective action by the air bag to stop the occupant's movement.

Assuming that the restraint is applied uniformly throughout the deceleration period, it takes 37 g's of deceleration to stop an object from 40 miles per hour within 17 inches. The air bags now being tested do not, however, work with this efficiency, because they do not apply a constant resistance; the inflated bag exerts only soft resistance during the first eight inches of the object's penetration into the bag and then rapidly becomes stiffer, like a spring "hardening" under an increased load. For this reason it is doubtful that the air-bag systems under trial would do an acceptable job in present full-size automobiles even in collisions at 35 or 40 miles per hour. And at higher speeds or in smaller cars the air bag certainly will not provide protection unless the deployment time is considerably reduced and the restraining efficiency is improved.



HONEYCOMB STRUCTURE made of heavy aluminum foil is the energy-absorbing part of the additional paneling installed in the interior of the passenger compartment. The aluminum structure is not visible to the passengers because it is covered with a standard paneling material. At top is a piece of the honeycomb material before it has received an impact, at bottom a piece that has been hit by a steel post designed to simulate part of a human body. In tests by the Calspan Corporation the aluminum material has proved to be somewhat too resistant, so that the automobile-crashworthiness group is investigating paper honeycombs.

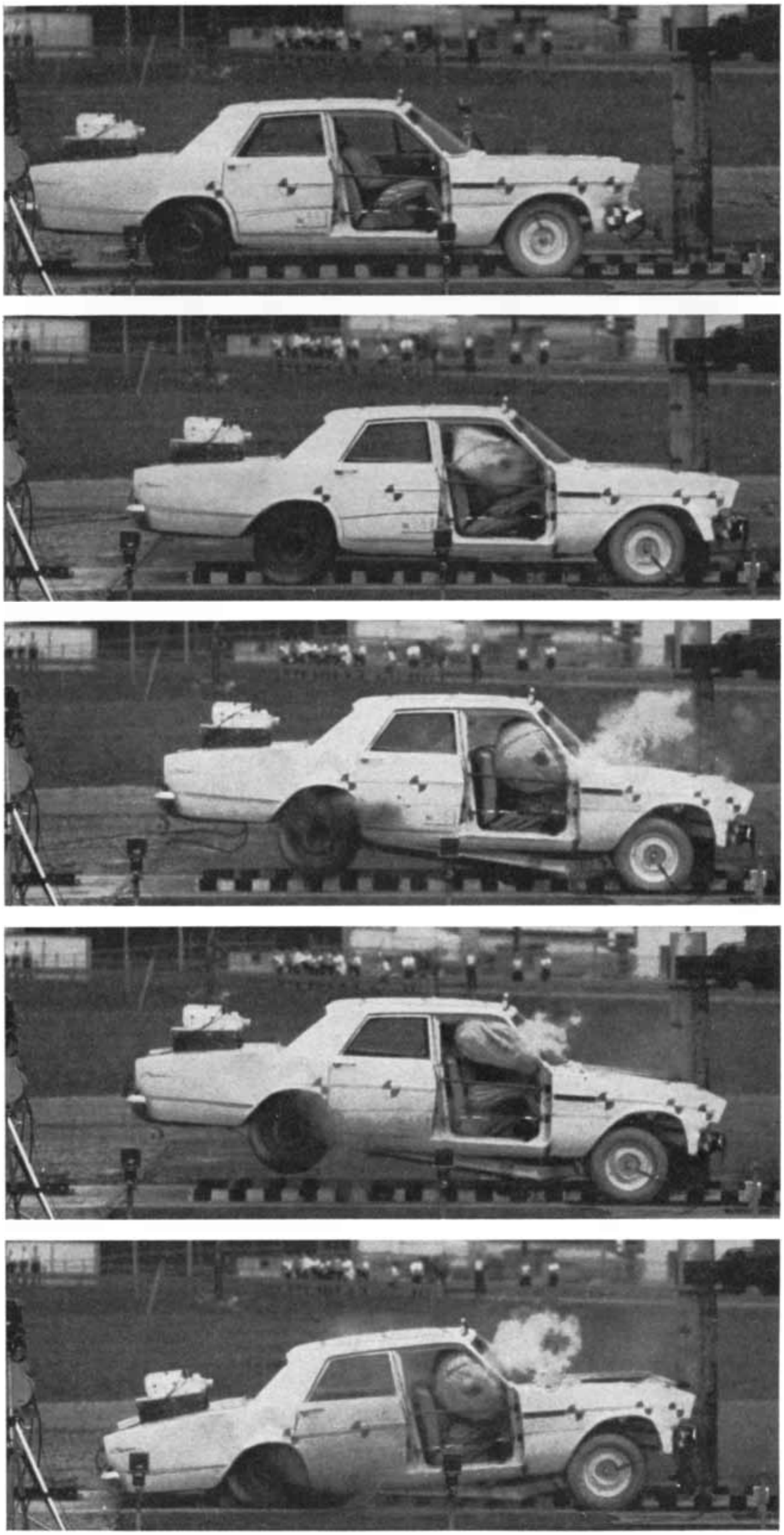
There are other reasons for judging that at present the air-bag approach does not look very promising. The system is extremely complicated and therefore would be rather expensive and not as reliable or immune to functional failure as one would like. In its present versions it offers protection only in frontal collisions, giving no relief for the important problems of side collisions and roll-overs. On the whole it appears that the air-bag concept as it has been developed so far cannot compare in effectiveness with belts and harnesses, which have proved themselves to be efficient devices for restraint.

Restraint systems in the "active" category offer more options in design and probably could provide the most satisfactory solutions to the crashworthiness problems, if all motorists could somehow be induced or required to use them. Efforts have been made to convert a lap-and-shoulder harness into a "passive" system—making motorists an offer they could not refuse. In this design a person entering an automobile would automatically be strapped in by the harness as he took his seat. The idea does not appear truly feasible, however, because the equipment would be cumbersome and the belts would have to be anchored to the doors, obviously an insecure location for them in some types of accident.

Investigation of methods for the protection of the occupants in side-collision and roll-over accidents is directed to three objectives: ameliorating the effects of impacts with the sides of the vehicle (since contacts with them cannot be avoided in such accidents), providing safeguards against ejection of the occupants and strengthening the door pillars and headers to maintain integrity of the roof in roll-overs.

Recent studies in our laboratory have indicated that for effective buffering of impacts with the lower sections of the doors they should be padded with three to four inches of a material with high energy-absorbing properties. It should be a rather firm material that yields to pressure at a constant rate as it is penetrated and recovers its form only slowly. Some of the foam polyurethane materials, in combination with crushable supporting structures, appear to be ideally suited for this purpose.

Because thick padding on other parts of the automobile's side structure would interfere with vision, these areas should be so constructed that they do not take the full impact of the part of the body thrown against them (for example the head) but shunt it onto a window or other absorbing area. A new type of glass for the side and rear windows is being investigated. The conventional glass for these windows is a heat-treated version that during collisions shatters into small granules so that the flying fragments do not cut the occupants. In order to give the windows high resistance to penetration and shattering, with a view to preventing ejection of the occupants, the glass now being studied for the side and rear windows is a form of sandwich such as is already in use for windshields. The present windshield glass has a thin layer



TEST OF AIR BAG is made at the Calspan track. The series of photographs is from a motion-picture film. The figure in the front seat is a dummy that is partly restrained by a seat belt of the lap type. As the car hits the steel pillar at right at a speed of 41 miles per hour, the dummy is first thrown forward into the air bag and then hurled back as the bag expands. Tests of this type have led the author to the view that at the present stage of development the air bag offers too little resistance in the first moments of a crash and too much later on.

of transparent, energy-absorbing polyvinyl butyral bonded between two layers of glass. In the version for the side and rear windows the glass forming the sandwich cover on the interior side of the car can be replaced by a polyester material to prevent the release of lacerating fragments in a collision.

We have been doing basic research on the external structure of the automobile by designing and testing modifications intended to cope more effectively with the fundamental problems of preventing intrusions into the occupants' "survival space" and dissipating the energy released in a collision with a minimum of harm to the occupants. The major attention has been given to frontal collisions and collisions involving the sides of the vehicle, including roll-overs.

In frontal collisions it appears that the conventional structure of present full-size automobiles is adequate to protect the occupants when the car collides with a fixed flat barrier at speeds under 40 miles per hour, provided that the occupants are adequately restrained. We have therefore focused mainly on structural modifications designed to improve protection in collisions at 40 to 60 miles per hour. At such speeds three to five feet of crushing compression of the vehicle's front-end structure would be required to spread out the deceleration of the occupants so that it did not rise above tolerable levels (perhaps no higher than 40 g's).

To allow a structural collapse of this length some provision must be made to prevent the engine from being rammed into the occupants' compartment. Two approaches are under study: lengthening the frontal, under-the-hood section of the vehicle and/or providing a ramp to shunt the engine under the passenger compartment when the engine is driven back.

For our first tests of modifications to make a vehicle more crashworthy at speeds above 40 miles per hour we placed a deflection ramp on the fire wall between the engine and passenger compartments and made structural changes in the frontal section of a conventional automobile that were designed to make the rate of deceleration as uniform as possible throughout its duration; for this purpose the absorption of energy by the forward part of the structure was increased and the absorption by the rear part was reduced. The performance of the modified automobile was then compared with that of a conventional one in a test in which they collided with a rigid pole at nearly 60 miles per hour [*see illustration on page 79*]. In the modified vehicle the engine, as had been hoped, was deflected under the passenger compartment instead of being pushed into it, and the deceleration of the compartment produced relatively little damage to its integrity. The peak deceleration of this compartment during the collision was reduced from 85 *g*'s in the conventional automobile to 60 *g*'s in the modified one.

Refinements are now being made to avoid too great aggressiveness of the modified vehicle in collision encounters with other automobiles. To this end the structural strength of the front end has been somewhat reduced and it has been equipped with a grille structure that distributes the impact loading more uniformly. The tests of these refinements so far indicate that the crashworthy automobile is no more aggressive than a conventional car in a collision with another full-size car. What the effects of a collision would be on a small car of compact or subcompact size has not yet been examined. It is well known that collisions between full-size (4,000-pound) and subcompact (2,000-pound) automobiles of conventional design generally have severe results for the occupants of the small vehicle. In view of the fact that about a fourth of the new automobiles now being bought in the U.S. are in the subcompact class, the possible effects on such cars of encounters with vehicles having the frontal modifications we are considering will have to be investigated carefully in relation to the overall vehicle population mix before the potential benefit of the modifications can be evaluated.

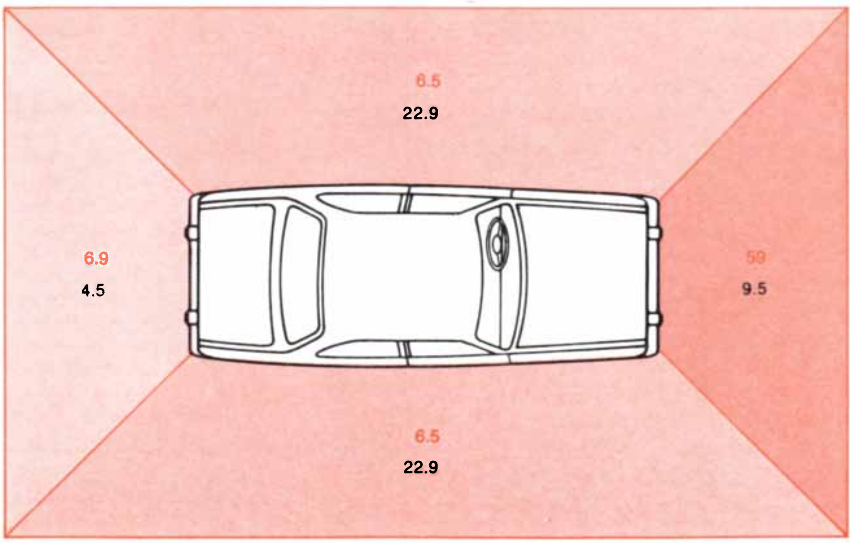
In addition to the frontal modifications, we have also been testing changes to strengthen and protect the sides of an automobile. These modifications include cross members between the side rails to strengthen the frame under the passenger compartment, crossbeams (placed in the doors) to buttress the side pillars and a roll bar across the roof to forestall a roof cave-in in roll-overs. For protection of the occupants from impacts, critical locations in the passenger compartment are padded and an energy-absorbing laminated glass (also designed to resist ejection) is used in the side windows. The efficacy of these modifications has been tested in collisions in which the vehicle was struck in the side (at a 90-degree or a 45-degree angle) by a full-size conventional automobile traveling at 45 miles per hour and also in side collisions in which the modified vehicle struck a fixed object with an impact speed of more than 20 miles per hour. The experimental vehicle proved to be substantially more successful than conventional cars are in resisting intrusions into the passenger compartment in all these tests. The window glass of the modified vehicle remained intact, whereas the unlaminated glass in a conventional automobile given the same impact was shattered. The deceleration responses of unrestrained dummies in the front seat on the struck side of the modified vehicle were so moderate that it is believed the deceleration effects even in severe collisions will be within tolerable limits for human passengers. It is not yet known how efficacious the side-strengthening modifications will be in roll-over accidents, as this has not yet been tested.

It appears that the tested modifications of the frontal and side structures will not present difficulties in the operation of the automobile, and that it can be manufactured by conventional methods. Consequently the research program at the Calspan Corporation has proceeded to the construction of prototype vehicles, incorporating the various protective modifications into a conventional full-size automobile. The strengthening features add about 10 percent (approximately 400 pounds) to the vehicle's weight.

The features embodied for crashworthiness include urethane foam behind the front bumper (to absorb energy in collisions below 10 miles per hour) and reinforcements that strengthen the frame against both frontal and side collisions [*see illustrations on pages 80 and 81*]. A grille in front is designed to spread out the impact forces in a frontal collision with another car, and the front inner fenders, along with the hood, help to absorb the impact. A deflecting panel of high-strength steel is placed on the fire wall to prevent the engine from being driven into the passenger compartment. The pillars and other structures on the sides and roof of this compartment are reinforced. The interior has padding on the door, pillar and header areas; the angle of the steering wheel is changed, and the side windows are of laminated glass and plastic.

These additions to the automobile structure, together with a suitable restraint system for the passengers that would always be used, could be expected to give reasonable protection to the occupants in most collisions. There remains, however, the problem of what the effect of this structure might be on smaller vehicles with which the strengthened full-size car collided. Clearly the reinforced automobile could not be recommended if it would become a serious menace to the millions of compact and subcompact cars now on the highways. The problem of protection for the small vehicle might be solved by applying protective measures, such as are described here, to the maximum degree in the 2,000-pound subcompact and designing all heavier vehicles so that they are compatible with the small vehicle in collisions.

Although this problem and others are still to be solved, the crashworthiness research program at least has already demonstrated that a significant improvement in passenger protection is possible.



ACCIDENT FREQUENCY in different areas of an automobile is depicted in terms of percent of all injury-producing accidents (*color*) and percent in each group that were regarded as severe (*black*). In addition to the categories shown, accidents in which a car rolls over account for 20.7 percent of the total, and 23.1 percent of them are severe. The data suggest what components of an automobile are most in need of improved crashworthiness.

Guts.

We lay it on the line. BMW builds the best sports sedans in the world.

Who says?

The drivers do. In poll after poll, men who know cars acknowledge BMW for what it is.

A car with an overhead-cam engine built around a combustion system so advanced that a BMW will cruise—*cruise*—at 100 mph.

A car with a suspension system that gives you roadholding qualities that are simply incredible. So efficient that many racing cars have imitated it.

A car that for all-out performance, for pure quality of construction, for right-down-to-it *guts*, has no equal.

Do you need only transportation? Then BMW is not for you.

There are always the so-called economy cars, or any number of less exciting choices.

But if you want the best sports sedan money can buy, if you want a car that makes driving a joy, then visit your BMW dealer.

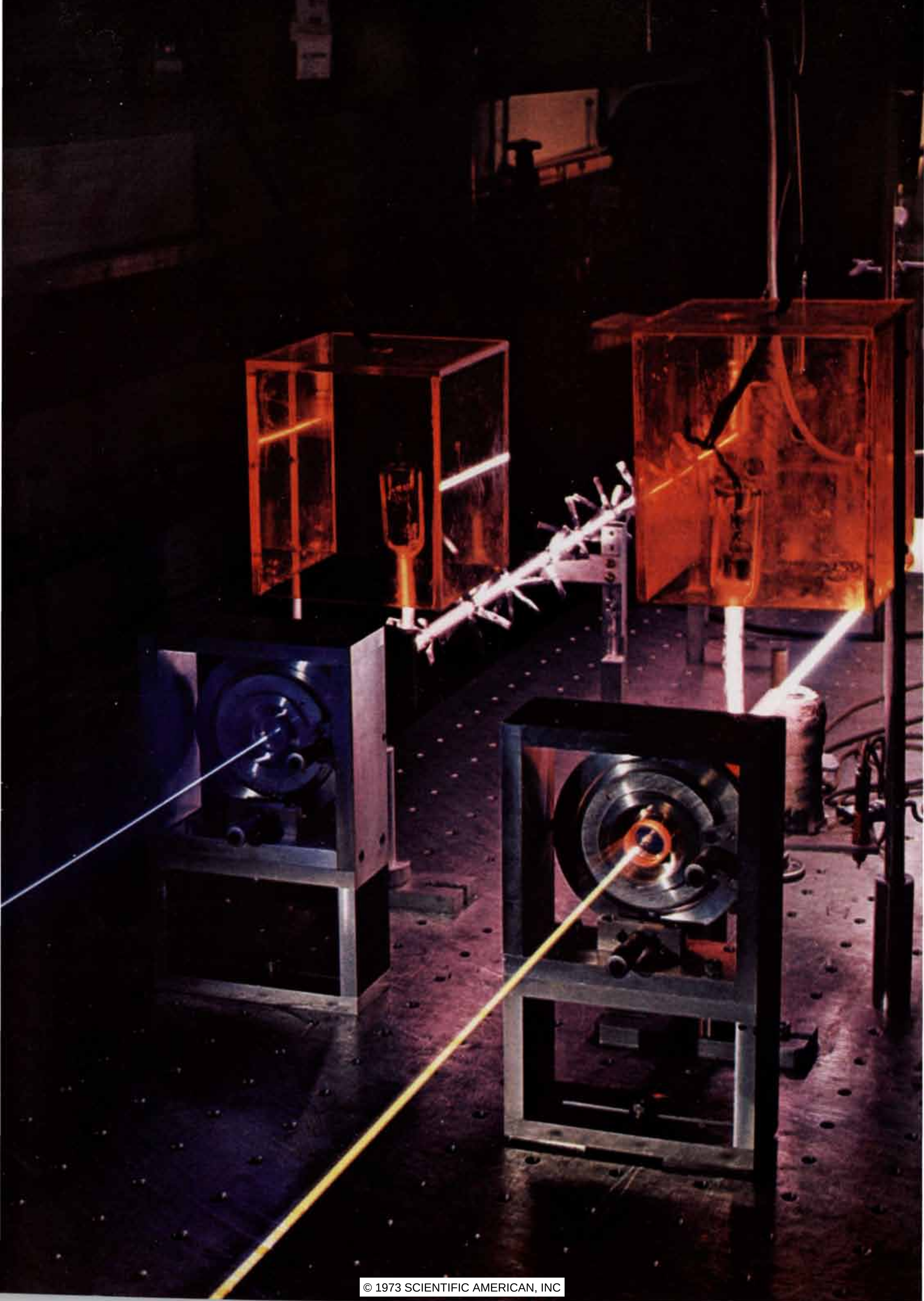
Has he got a car for you!

BMW's are available in four or six-cylinder models. U.S. Sole Importer: **Hoffman Motors Corporation.** 375 Park Avenue, **New York,** N.Y. 10022, 12541 Beatrice Street, **Los Angeles,** California 90066. Dealers Coast to Coast. Advantageous European Delivery Program. Write for free brochure.

Canada: East: BMW Distributors Eastern Canada, Ltd. 735 Progress Avenue, Scarborough, Ontario. West: BMW Distributors (Western) Co. 2040 Burrard Street, Vancouver 9, B.C.

Bavarian Motor Works

BMW



METAL-VAPOR LASERS

Among the many systems that can produce laser light is a vaporized metal mixed with another gas. Such lasers generate continuous laser radiation at good power levels and over a wide range of wavelengths

by William T. Silfvast

Since 1960, when a laser was first successfully demonstrated, laser action has been obtained in many different physical systems. The original ruby laser was succeeded by other solid-state lasers, including the neodymium laser and the large class of semiconductor lasers. Among liquid lasers the dye laser has attracted the most notice. Gas lasers, which appeared soon after the first ruby laser, have been extended to such special types as TEA lasers (TEA stands for transverse electrical excitation at atmospheric pressure), chemical lasers and gas-dynamic lasers. Currently there is considerable interest in another special class of gas lasers: metal-vapor lasers. In a variety of metal-vapor systems it has been possible to produce continuous laser radiation of many different wavelengths in the visible, infrared and ultraviolet regions of the spectrum. Moreover, metal-vapor lasers are relatively simple to build and to operate.

"Metal" normally suggests a solid material, whereas "vapor" suggests a gas. Hence the term "metal vapor" might at first seem to be contradictory. The reader will recall, however, that any chemical element can exist as a solid, a liquid or a gas, depending on the temperature and pressure of its environment. When "vapor" is used to describe a gaseous state, it is implied that the gas is in the presence of its liquid or solid state. A familiar example is the coexistence of water vapor (steam) and liquid water in a vessel on the kitchen stove. So it is with metal-vapor lasers.

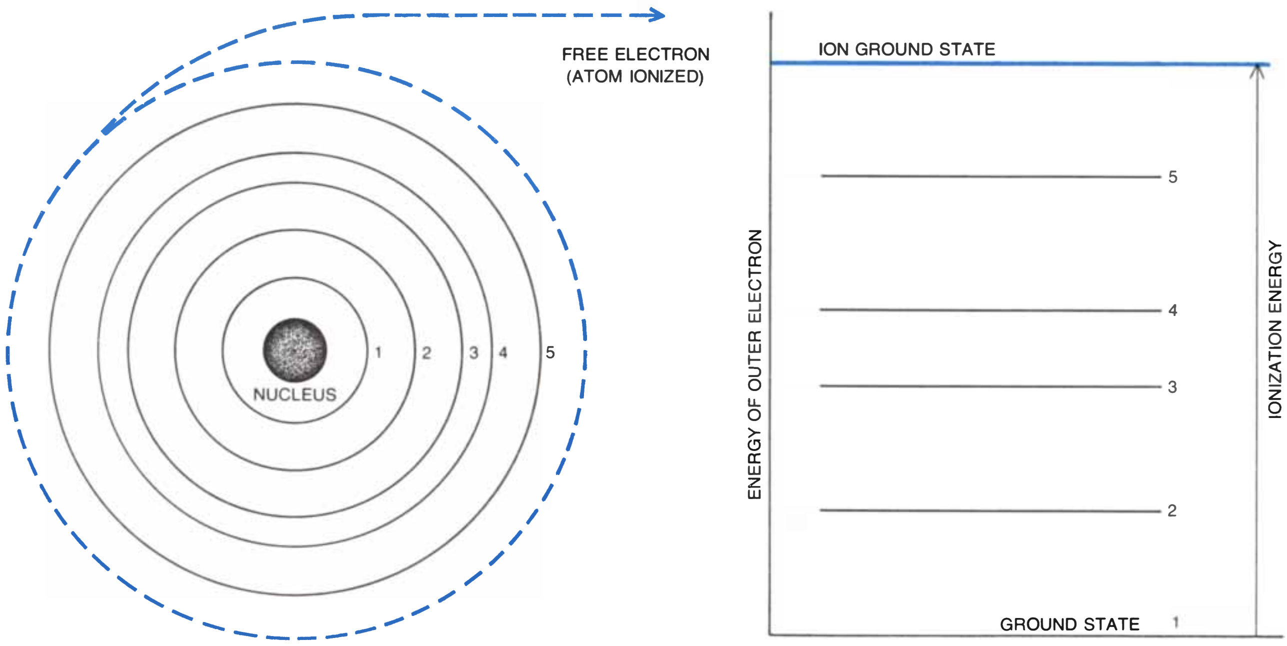
The first such laser was a cesium-vapor laser developed by S. Jacobs, P. Rabinowitz and Gordon Gould at Columbia University in 1961. They supplied energy to the vapor by means of a helium flash lamp. The laser's output was in the infrared. Some two years later J. Rigden and A. H. White operated the first mercury-vapor laser at the Bell Telephone Laboratories. Other early work included the efforts of W. E. Bell and Arnold Bloom of Spectra-Physics, Inc.; of G. R. Fowles, R. C. Jensen and myself at the University of Utah; of William Walter, Martin Piltch and their colleagues at TRG, Inc., and of A. Tibiloz in the U.S.S.R. Most of this work was directed toward achieving laser action by passing very strong pulses of electric current through a rarefied gas that also contained a metal vapor.

The early efforts produced pulsed laser light in the vapors of cadmium, zinc, tin, lead, manganese, copper, calcium, germanium, indium and gold. Because the pulses of laser light were brief (10^{-8} to 10^{-5} second) and the efficiency of the lasers was rather low the early metal-vapor lasers did not compare favorably with other types of laser for experimental or commercial applications. The helium-neon laser, which emits red and infrared radiation, and the argon-ion laser, which emits blue and green light, attracted most of the research and development on gas lasers because they both provided a continuous output of laser light, as opposed to the pulsed output of the metal vapors. Another consideration was that the helium-neon and argon-ion lasers used pure gases, which are easier to work with than vapors. Moreover, the helium-neon laser promised a low cost and the argon-ion laser offered a continuous output at high power (up to several watts).

The effort to develop metal-vapor lasers never really gained momentum until Fowles and B. D. Hopkins obtained a quasicontinuous output of laser light in a helium-cadmium alternating-current discharge. (The term "discharge" in this context describes an electric current flowing in a low-pressure gas.) Their finding was subsequently confirmed by our own work at Bell Laboratories: we demonstrated the first direct-current, continuous-wave output of laser light in a metal vapor. Our measurements indicated that the helium-cadmium laser was capable of high power and efficiency. Development efforts on the helium-cadmium laser were also pursued by John P. Goldsborough at Spectra-Physics and J. R. Fendley and K. G. Hernqvist at the RCA Laboratories. The discovery of a new laser line in the ultraviolet region of the spectrum, where laser output had been rare, made the helium-cadmium system even more attractive.

As a result of these developments the concept of metal-vapor lasers acquired a new interest. Not long ago Marvin B. Klein and I developed another metal-vapor laser that we believe will become a significant commercial instrument. This laser employs selenium vapor and operates continuously in a helium-selenium discharge. It produces many different colors of laser light, including blue, green, yellow, orange, red and infrared. Both the helium-cadmium laser and the helium-selenium laser are rela-

TWO LASERS operating on the metal-vapor principle appear in the photograph on the opposite page. The one emitting blue light is a helium-cadmium laser and the one emitting yellow light is a helium-selenium laser. Helium is energized by an electric current. Some of the stored energy of the helium is transferred to atoms of vaporized cadmium or selenium, which emit light as they drop to lower energy levels, thus providing the laser output.

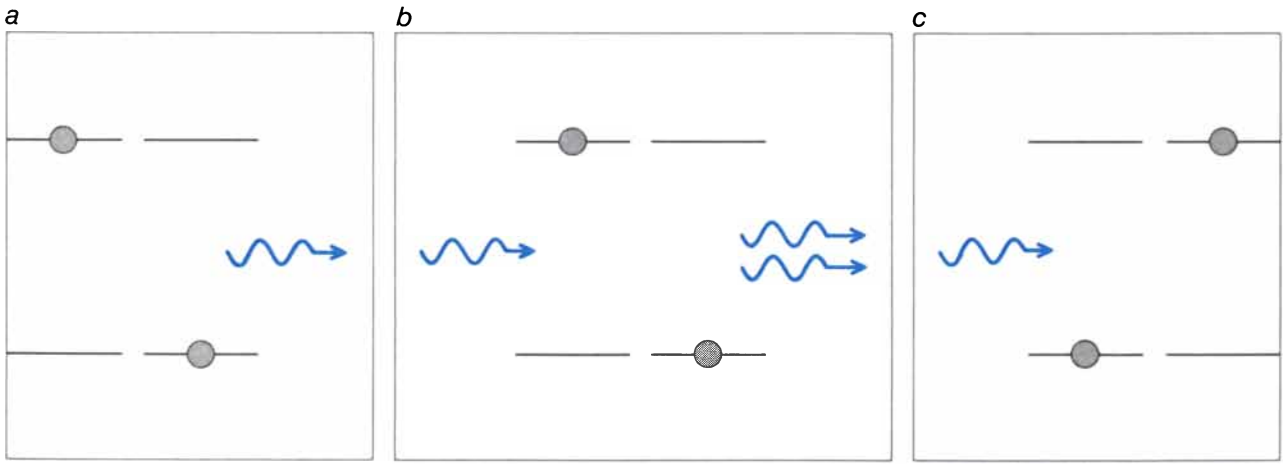


ORBITS AND ENERGY STATES that the outer electron of an atom might have as it circulates around the nucleus are portrayed. At left are the orbits, at right the corresponding energy states. Higher numbers indicate higher energy. Energy required to move an electron to a higher orbit or energy state or to detach it from the atom (*color*), thus ionizing the atom, must come from such outside events as collision with a fast-moving electron. Energized atoms radiate the photons, or quanta of light, that a laser amplifies.

tively simple to construct and operate. Many applications that seem suited for the particular characteristics of these lasers are evolving.

Readers of this magazine's earlier articles on lasers will recall that "laser" stands for light amplification by stimulated emission of radiation. In the continuous-wave metal-vapor lasers that we have been studying the key to laser action is the storage of energy in helium atoms. This energy originates with the electric current that produces an electrical discharge in a gas consisting of helium atoms and metal atoms. The electrons of the current yield some of their energy to the helium gas, where the energy is stored in various excited states of the helium atoms. A portion of this energy is later transferred from the helium atoms to the metal atoms during random collisions in the gas mixture. The metal atoms can be ionized (made to lose one or more electrons) in the collision process, and many are left in states where they can give up energy in the form of laser light. It will become apparent why metal vapors are particularly useful for these processes.

The process of stimulated emission of radiation was first described theoretically by Albert Einstein in 1917. It is a



EMISSION AND ABSORPTION of energy by an electron are depicted. In spontaneous emission (*a*) an excited electron (*left*) falls to a lower energy state, giving up energy. The yielded energy is emitted in the form of a photon (*color*) of an energy and therefore a wavelength equivalent to the energy given up by the electron. In stimulated emission (*b*), which is the process that takes place in a laser, a photon having the exact difference in energy between two energy levels of an electron collides with the atom, causing an electron in the upper level to radiate its energy as a photon, so that there are now two photons instead of one. Absorption (*c*) takes place if the electron is in the lower state. The electron absorbs the colliding photon and is raised to a higher energy state. Laser action requires a larger population of electrons at upper energy levels than at lower ones to overcome the absorption.

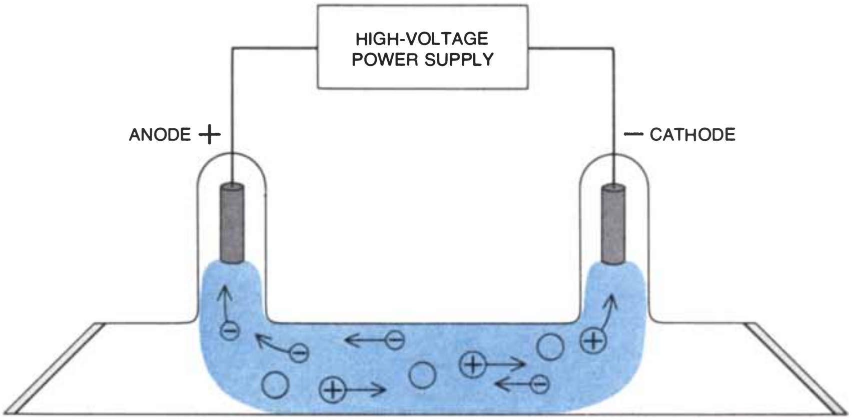
process whereby a light beam of a specific color or wavelength can interact with an atom in such a way that the atom will emit additional light that moves in the same direction and has the same wavelength as the original beam. If a beam were fed into a suitable laser amplifier, the amplifier would preserve the original direction and wavelength of the input beam and the intensity of the light would increase. An oscillation can be produced by placing mirrors at the ends of the amplifier, forcing the light to bounce back and forth between the mirrors. In such an oscillator the intensity of the light can build up to many orders of magnitude greater than the intensity of the original light signal, producing a highly directional laser beam of an extremely pure color. The light is extracted from the laser through one of the mirrors, which is designed to transmit a small percentage of the light.

The ability of atoms to store energy has to do with the electrons within the individual atoms. The electrons exist as a cloud of negative charge around the positively charged nucleus. Each electron occupies a state of energy and angular momentum that cannot be occupied by any other electron. Hence the electrons tend to fill stable shells surrounding the nucleus. The electrons of the outermost shell are the ones most easily affected by outside forces because of their accessibility. These outer electrons can be moved to higher energy states, but they always tend to return to their lowest energy state: the ground state [*see top illustration on opposite page*].

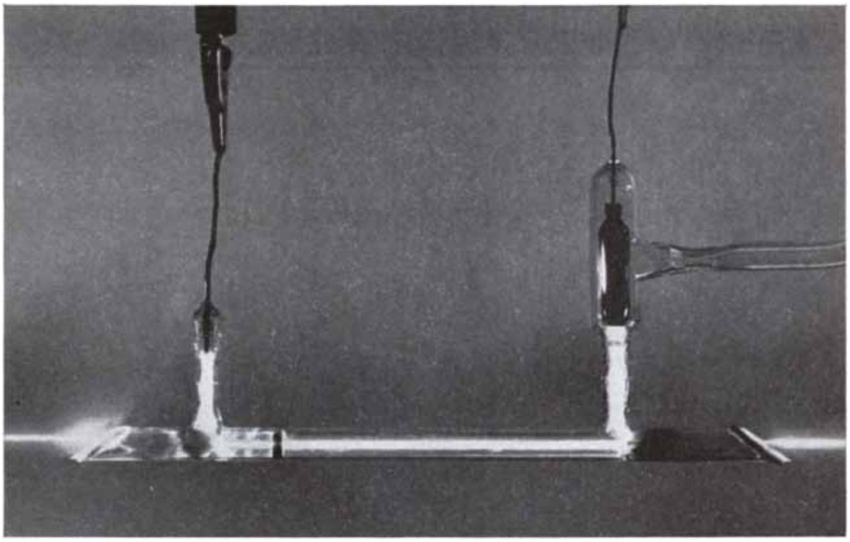
As an electron is moved to higher and higher energy levels, it will finally have enough energy to free itself from the attracting force of the nucleus, and it will escape from the atom. The electron then becomes a free electron with a negative charge and the atom becomes an ion with a positive charge. The energy necessary to remove an electron (in the ground state) from the atom is called the ionization energy. Each kind of atom has a different ionization energy depending on the number of protons, neutrons and electrons it has. The energy levels of the electronic states of atoms can be measured in terms of electron volts (eV): one electron volt is the amount of energy attained by an electron when it is accelerated through a potential difference of one volt. Helium has the largest ionization energy of any atom (24.6 eV) and cesium has the smallest (3.9 eV) [*see illustration on next page*].

When an electron is in one of the higher energy states, it will always seek a lower energy, sometimes by jumping directly to the ground state and sometimes by cascading down from one level to the next. In such transitions energy must be conserved, and so the atom gives off a photon, or quantum of light, that has an energy exactly corresponding to the difference between the two energy states of the electron. The frequency or wavelength of the light given off is related to the energy difference. A red photon has an energy of about 2 eV and a blue photon an energy of about 3 eV.

Electrons at certain levels decay (fall to a lower state) more easily than electrons at other levels. Each excited electronic state of the atom has a characteristic lifetime that indicates the average time it takes an electron to fall to a lower level and thus radiate a photon.



GAS DISCHARGE provides a means of exciting electrons. If a gas such as helium is put into a glass vessel at low pressure and a high voltage is placed between two electrodes, an electric field (*color*) is produced, applying a force to the charged particles and causing a current to flow between the electrodes. The electrons, having very little mass, are accelerated to extremely high velocities. In a metal-vapor laser they collide with helium atoms, raising outer-shell electrons of those atoms to a higher energy state. This energy is then transferred by collision to neutral atoms (*open circles*), thereby providing the basis for laser action.



GAS-DISCHARGE REGION of a helium-cadmium laser in the author's laboratory is the bright area between electrodes. The anode is at left, the cathode at right. The cadmium that is being vaporized to provide the metal-vapor atoms for the system is the slender dark object in the tube, a short distance to the right of the anode. Laser beam appears at left and right of slanted windows at the ends of the tube. This small laser, where the distance between windows is only eight inches, produces a continuous output of two milliwatts of blue light at a wavelength of 4,416 angstroms. The light is highly intensified and extremely pure.

Most excited states have lifetimes of about 10^{-8} second.

There are some excited states or levels in all atoms in which an electron cannot decay easily by giving up a photon. Such atoms must therefore wait for other means of giving up their energy, such as colliding with other atoms or with the walls of the system. Electrons in energy states of this kind tend to stay there for relatively long periods of time (.001 second or more) and are referred to as being in metastable states. They play an important role in storing energy, which can then be employed in the excitation process of a metal-vapor laser.

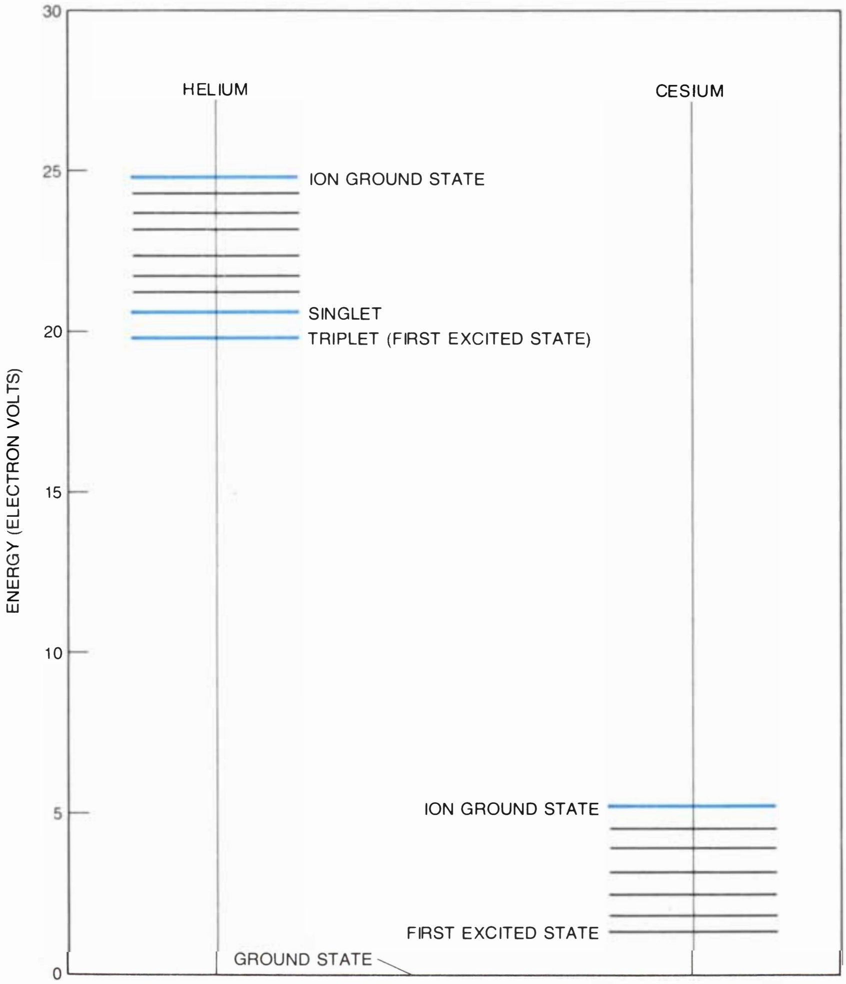
The normal radiative decay from a higher electronic state to a lower one is termed spontaneous emission. Processes also exist that can force an atomic electron to a higher state or stimulate it to jump to a lower state. An example of forcing is provided when a photon collides with an atom and excites the outer electron to a higher level, which can happen when the energy or wavelength of the photon corresponds exactly to the difference in energy between the state the electron is in and some higher possible state. This process is known as absorption because the photon is actually absorbed by the atom and all the photon's energy goes into raising the electron to a higher state.

Similarly, the stimulated electron can move to a lower level, provided that such a level exists and that the difference between the two levels corresponds exactly to the energy of the photon. This is the process of stimulated emission. The energy given up by the electron in jumping to a lower state goes into creating an additional photon with the same characteristics as the colliding photon [*see bottom illustration on page 90*].

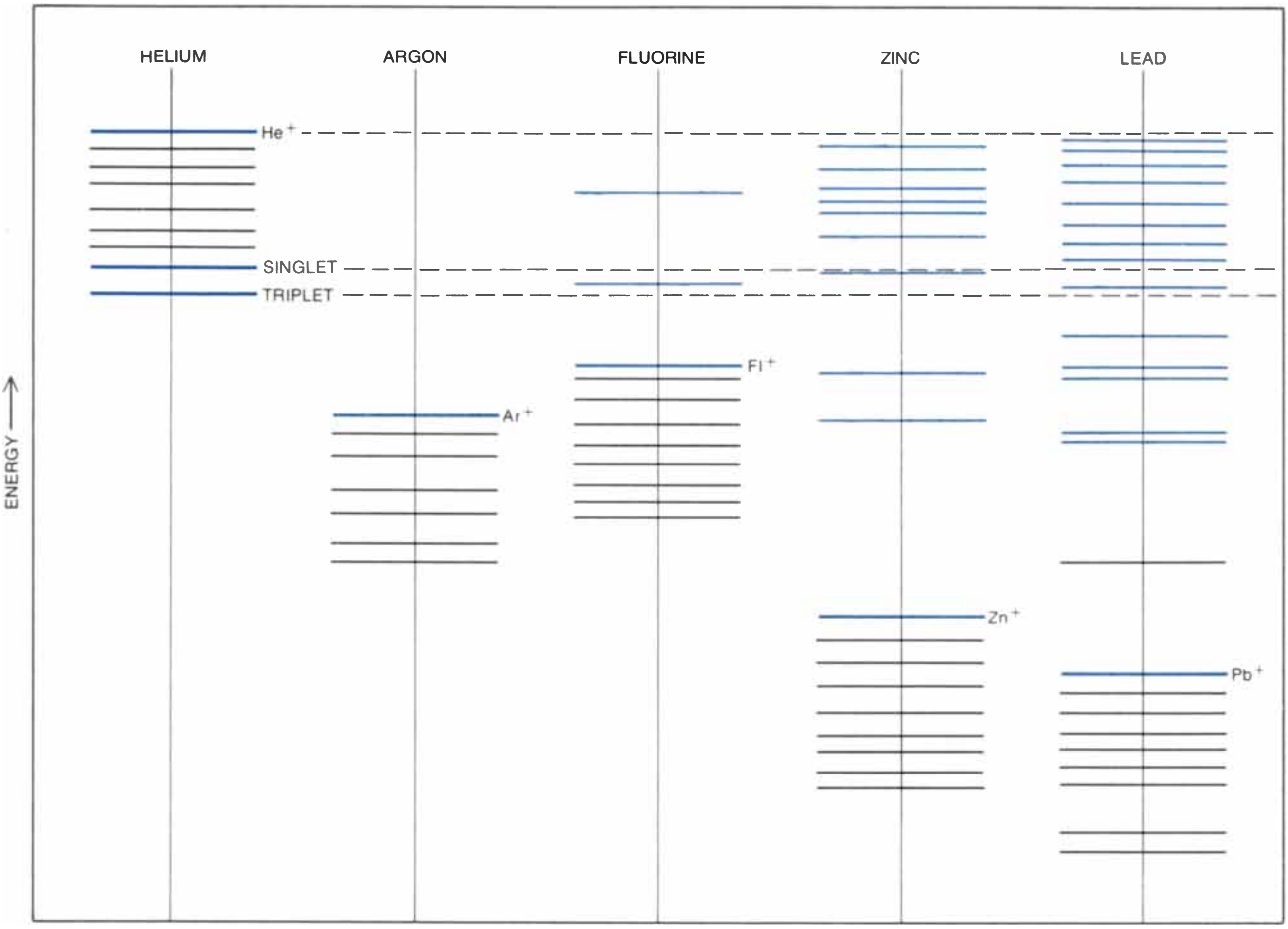
A large number of atoms can provide an increase in light in a laser amplifier if the population of electrons in the excited states of the atoms is suitably arranged. Consider two excited levels of a system of identical atoms with the electrons divided between the upper and lower levels. If a light beam having a wavelength corresponding to the difference in energy between the two levels is allowed to pass through the medium, it will be amplified if there are more atoms with electrons in the upper state and absorbed if there are more atoms with electrons in the lower state. The condition of having more atoms in the upper state is called a population inversion (because it goes against the normal processes of nature, which tend to keep more electrons at lower energies than at higher energies). The search for new laser systems is therefore not easy, since one must find a means of working against the natural tendencies of the electrons in order to obtain more atoms with electrons in the upper level than atoms with electrons in the lower level.

One of the best ways to feed energy into a gas in order to obtain a population inversion is by means of a gas discharge [*see top illustration on preceding page*]. A gas discharge begins when a high voltage is established between two electrodes inside a glass vessel that is filled with a gas at low pressure. The voltage produces an electric field in the gas, and the field causes an electric current to flow between the electrodes by applying a force to the charged particles. The negatively charged electrons are accelerated toward the anode, and positively charged ions are accelerated toward the cathode. The electrons, which are much lighter than the ions, are accelerated to very high velocities (10^8 centimeters per second), and they lose part of their energy by colliding with atoms of the gas. Some of the electrons gain enough energy to ionize other atoms, creating more free electrons, which help to sustain the discharge.

There are two kinds of collision between the electrons and the atoms. One is elastic, resembling the impact of two billiard balls: the light electron bumps into the heavier atom, giving it a slight shove. In this type of collision the electron loses only a small fraction of its energy because of its small mass. In the other type of collision, which is inelastic, the electron affects one of the outer-shell electrons of the atom by moving it to a higher state. Here the free electron



POSSIBLE ENERGY STATES for the outer electron of an atom of helium and an atom of cesium are compared. Because of the amount of energy required to raise helium to the first excited state, which is much higher than the first excited state of cesium, helium is useful for storing energy in substantial quantities. Moreover, several states (*color*) are metastable, that is, they have relatively long lifetimes during which they can retain the energy.



STORAGE LEVELS of helium, which are the metastable levels capable of delivering stored energy to other atoms, are compared with the energy levels of several other atomic systems. Because of restrictions in the energy-transfer process during collision, only the ion levels of other atoms can receive energy from the helium metastable levels, except in the case of neon. Argon and fluorine have no groupings of ion levels that are below the helium metastable levels and that also have other properties required for laser action. Zinc and lead, which represent two metal-vapor systems, have many ion levels below the metastable states of helium.

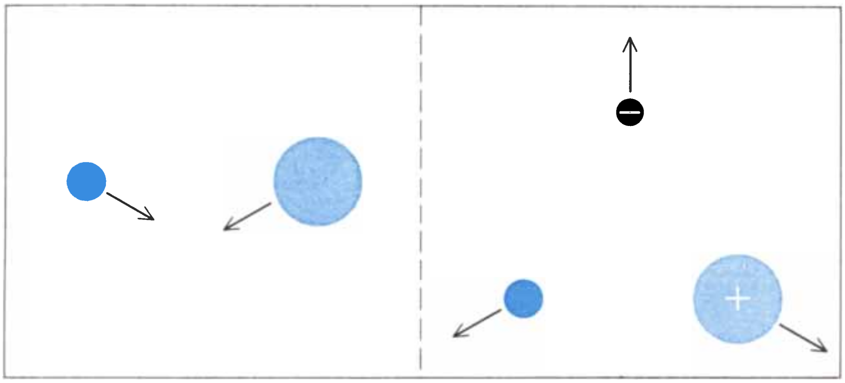
can lose a large fraction of its energy because many of the atomic states are separated by energies of several electron volts.

Helium, among all the atoms, has the largest energy gap between the ground state and the first excited state [*see illustration on opposite page*]. Thus when an electric current is induced in helium to form a gas discharge, the free electrons must be accelerated to an energy higher than 20 eV before they can take part in inelastic collisions with helium atoms. At the opposite extreme, however, the free electrons of a discharge in pure cesium need an energy of only slightly more than one eV to take part in inelastic collisions with cesium atoms. Thus the distinctively different energy-level structures in helium and cesium help to produce a high average energy of electrons in a helium discharge and a low average energy in a cesium discharge.

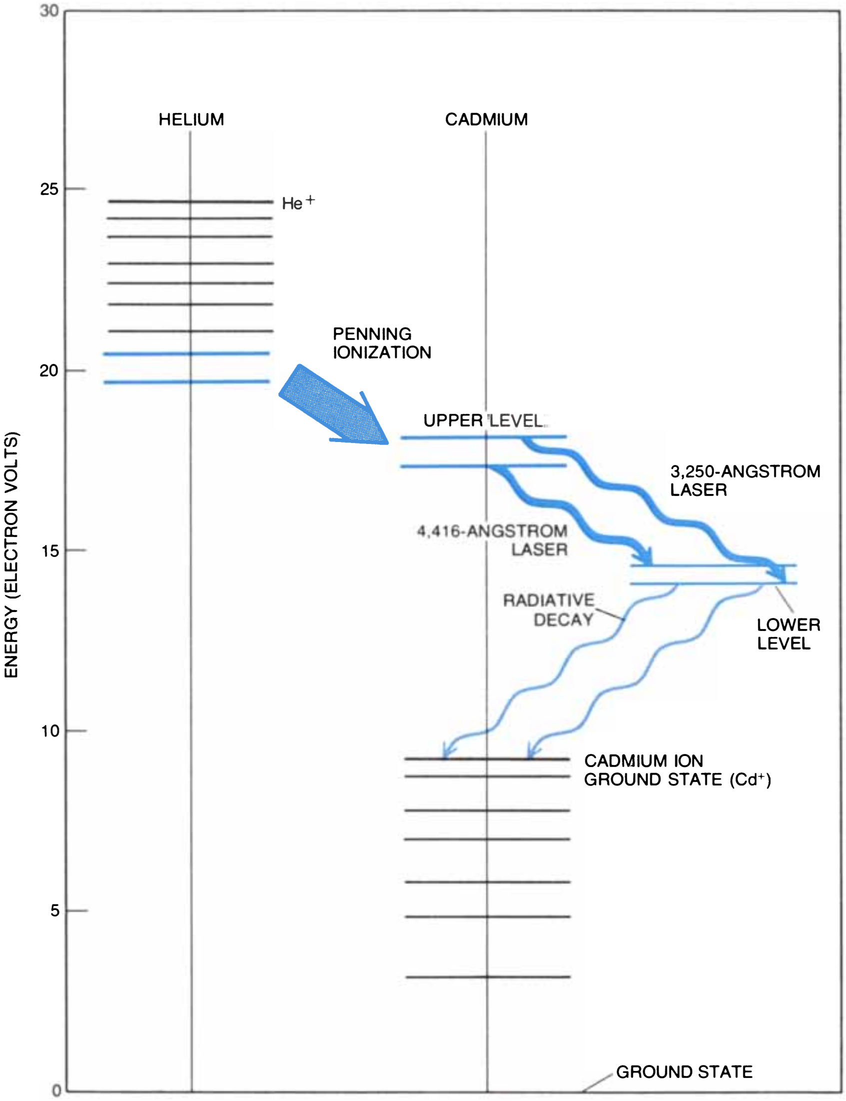
The high electron energy in the helium discharge causes many helium atoms to be excited to the very energetic excited states. Three of these states are metastable, that is, they have long lifetimes. They are respectively termed the singlet metastable state, the triplet metastable state and the ion ground state. The singlet and triplet metastable states have long lifetimes because they have no means of radiating their energy. The ion ground state, which is the lowest energy state of the ionized species (one electron removed from the atom), has a long lifetime because it must capture an electron before it can radiate its energy. Many of the excited helium atoms tend to end up trapped in one or another of the three metastable states, creating a large reservoir of electronic energy in the discharge.

If other atoms, which can be referred to as impurity atoms if their concentration is relatively small, are introduced into a helium discharge, the random collisions in the discharge will mix the impurity atoms with the helium atoms, and quite frequently the two types of atom will collide. During such collisions the energy stored in the metastable helium atoms can be transferred to the impurity atoms, leaving the impurity atoms in an excited electronic state from which they can subsequently emit radiation as they fall to lower levels. It is possible to obtain a population inversion between two excited levels of the impurity atoms provided that certain requirements are met.

One requirement is that the impurity atom must have suitable ion energy levels. To be suitable they must have an energy equal to or lower than the energy of the helium metastable levels, since the metastable helium atoms cannot give up any more energy than they have to begin with. Only the ion levels of other atoms can receive energy from the helium metastable levels because of restrictions governing the energy-exchange



ENERGY TRANSFER from a metastable helium atom (*dark color*) to a cadmium atom in a helium-cadmium laser involves a collision. At left the atoms are shown just before the collision, at right just after it. In the transfer of energy the helium atom returns to its ground state and the cadmium atom is ionized. Energy difference between helium metastable state and cadmium ionization state is carried away by an electron. Process is known as a Penning ionization after the late F. M. Penning of the Philips Research Laboratories.



PENNING IONIZATION in a helium-cadmium laser excites cadmium atoms to ionized states. Radiation from the upper levels of cadmium ions initiates blue laser action at 4,416 angstroms and ultraviolet laser action at 3,250 angstroms. Atoms decay from the lower levels faster than from higher ones, so that the required population inversion of excited ions is maintained. Cadmium levels above laser levels exist but are not relevant to this laser.

process during the collision. Neon is the only exception, the reason being that it has a high ionization potential. (The helium-neon laser involves an energy exchange from helium metastable levels to neutral helium levels.)

A second factor concerns the lifetimes of the ion energy levels of the impurity atoms involved in the population inversion. The upper level of a two-level system with a population inversion radiates energy corresponding exactly to the difference between the two levels. The electrons that have thus yielded energy by radiation end up at the lower level. If they remained there, a situation would soon arise wherein there would be more atoms at the lower level than at the upper level, and there would be absorption instead of gain. For continuous laser operation one must therefore find impurity atoms with combinations of levels such that the lower level has a short lifetime and the upper level is preferentially populated. This necessity rules out the ion ground state as a possible laser level, since all ion ground states are metastable. The pair of ion laser levels must be energetically above the ion ground state but still below the metastable helium states from which they receive energy.

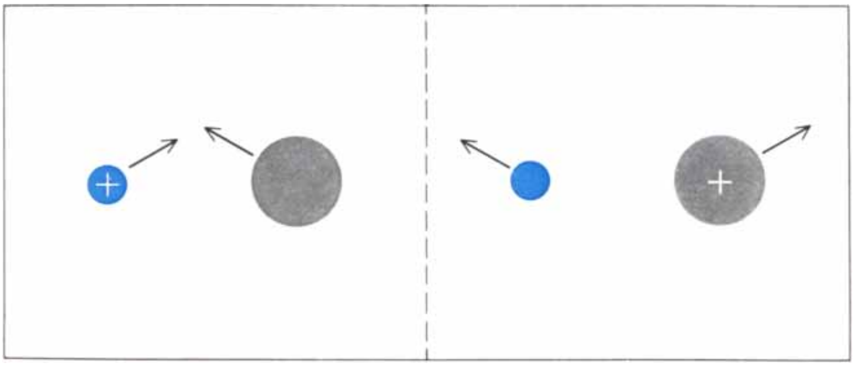
A final requirement is that the impurity atoms be vaporized. As a vapor they are in a gaseous state, so that their energy levels represent the discrete levels of single atoms or molecules as opposed to the smeared-out energy levels of a liquid or a solid. Some elements, which require a temperature of 2,000 degrees Celsius or higher to be significantly vaporized, would not be practical for vapor lasers.

With these requirements in mind one can look at the periodic table of the elements to see which ones would make good candidates for lasers. It turns out that a number of metal atoms meet the requirements. The possibilities can be illustrated by a comparison of the energy states of helium with those of two nonmetallic elements (argon and fluorine) and two metals (lead and zinc).

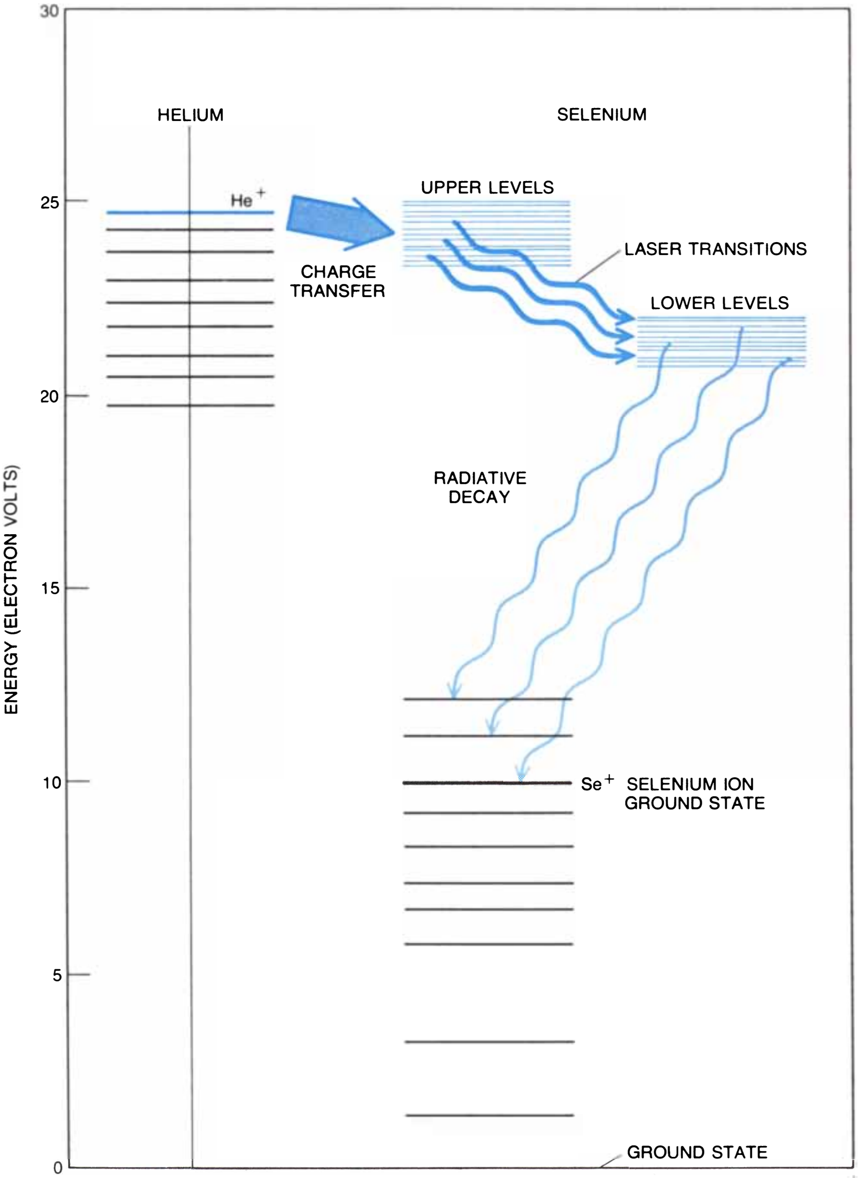
In the case of argon only the ion ground state can be reached from the metastable levels of helium, which clearly eliminates argon as a candidate for this kind of laser [*see illustration on preceding page*]. Fluorine has two levels above the ion ground state that could receive energy from the metastable levels of helium, but they are themselves metastable levels and so do not provide a basis for efficient radiation of energy. Lead and zinc, which are representative of typical metal atoms, have low ionization energies and also a considerable number of levels between the ionization energy and the metastable levels of helium that are possibilities for laser action. To be fully satisfactory, of course, the energy levels of a metal must also have appropriate lifetimes and good transfer rates in order to obtain an efficient transfer of energy from the helium.

Let us now look more closely at the two metal-vapor lasers I described briefly at the beginning of this article: the helium-cadmium laser and the helium-selenium laser. The helium-cadmium laser utilizes the triplet and singlet metastable levels of helium [*see bottom illustration on opposite page*]. In the energy-transfer process the metastable atoms of helium collide with neutral atoms of cadmium that are in their ground state, and the metastable energy of the helium is given to the cadmium. In order to conserve energy and momentum the cadmium is ionized, thus emitting a free electron to the discharge. With three separate particles, the only way the balance of energy and momentum can be achieved is for the electron to carry the exact difference in energy between the cadmium energy level and the metastable level of the helium atom. This type of collision and ionization is known as a Penning ionization after the late F. M. Penning of the Philips Research Laboratories, who first recognized the process.

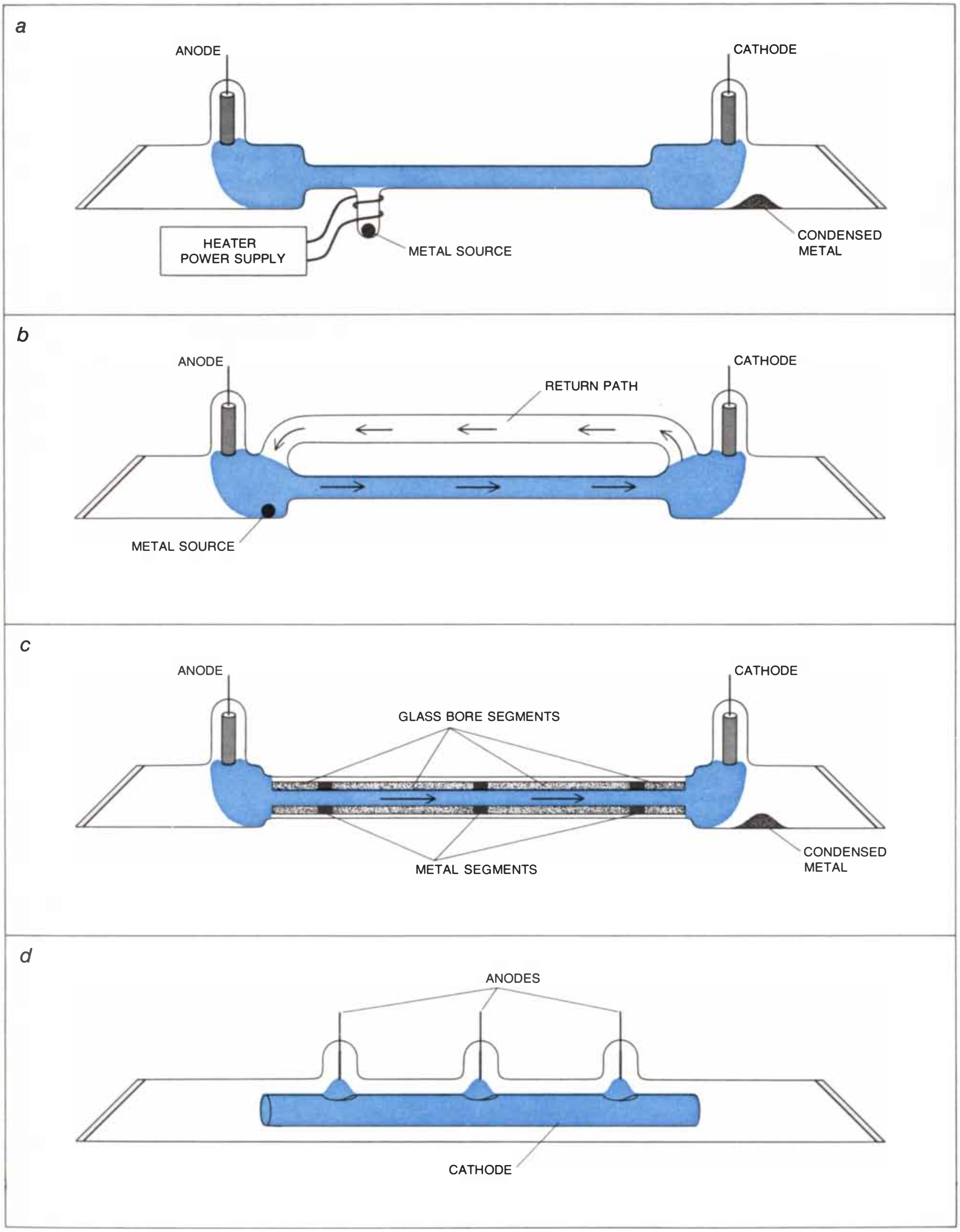
The cadmium system easily satisfies the lifetime requirements I have been discussing. The upper laser levels have lifetimes of about .5 microsecond, whereas the lifetimes at the lower levels are only a few nanoseconds—some 230 times shorter. Moreover, from measurements made of excitation rates to the upper and lower levels it appears that the transfer of energy from helium to the upper level proceeds at a rate about three times greater than the rate of transfer to the lower levels. Another attractive feature of the system is that the cadmium needs to be heated to only about 250 degrees C. to provide a sufficient density of cadmium vapor for the transfer of energy to proceed at a large enough rate for laser action. Because of the relatively efficient excitation process and also the favorable lifetimes of the upper and lower laser levels, the helium-cadmium laser has produced as much as 300 milliwatts of laser power at a blue wavelength of 4,416 angstroms and 50



CHARGE-TRANSFER IONIZATION in a helium-selenium laser involves a collision between an ionized helium atom (*color*) and a neutral selenium atom (*gray*). An electron from the selenium transfers to the helium, leaving helium neutral and selenium ionized.



LASER ACTION in a helium-selenium discharge arises when a selenium atom receives energy from a helium ion by giving up one of its electrons and raising another outer-shell electron to one of 13 upper laser levels. From there the energy is radiated to one of 11 lower levels, initiating laser action. Lasing has been produced from 46 such transitions.



LASER TUBES designed for continuously operating metal-vapor lasers are depicted. In each case the colored area indicates the discharge region. One tube (*a*) employs a flowing process termed cataphoresis pumping. The electric field of the gas discharge distributes the metal vapor. A diffusion return tube (*b*) recycles the vapor continuously. The simplified drawing omits a heater, separate discharges needed to keep the metal vapor from condensing in the end-window region and a grid that blocks the discharge from the return path. In the segmented-bore tube (*c*) metal segments are sandwiched between glass segments and heat from the discharge current vaporizes the metal. Hollow-cathode tube (*d*) has a long cylindrical cathode that heats the metal and excites the metal atoms.

milliwatts in the ultraviolet at 3,250 angstroms. Other continuous lasers that use Penning collisions to provide the energy transfer from helium have been demonstrated in zinc, tin and lead.

The helium-selenium laser derives its energy from the ground state of the metastable helium ions. The helium ions collide with neutral selenium atoms that have been dissociated from their molecular vapor form by collisions with electrons in the discharge, and since the helium ions have one electron missing they can readily accept an electron from selenium. The energy is transferred to the selenium atoms, leaving them in an ionized state, and the helium ions become neutral helium atoms in their ground state.

This type of reaction is known as a charge-transfer or a charge-exchange process because the electronic charge is transferred from selenium to helium. Since an electron is not released in the collision process, as it is in a Penning ionization, the only levels of selenium that can accept energy efficiently are the ones near the energy of the helium ions. Selenium has 13 levels that are close enough in energy to the helium ion ground state to receive energy and serve as upper laser levels [*see bottom illustration on page 95*]. These levels radiate light and fall to 11 lower states that serve as lower laser levels. The lower levels all have a fast escape route to the ion ground state to provide the necessary drain for a population inversion.

The helium-selenium laser, with its many wavelengths and relatively high efficiency, has produced a combined power of up to 250 milliwatts on six wavelengths of blue and green light; powers at individual wavelengths have been as high as 50 milliwatts. In addition, 19 wavelengths covering most of the visible spectrum have been achieved simultaneously from a single laser. Other metal vapors that have produced a continuous laser output by means of the charge-transfer process include cadmium, zinc and tellurium.

Since metal vapors exist in conjunction with either the solid or the liquid state of the metal, special laser tubes are required to contain the solid or liquid metal and to distribute the vapor within the gas discharge. In addition, a means of heating the metal to the appropriate vapor pressure is usually needed. Several types of gas-discharge tube have been designed to meet these requirements; four types are currently in use [*see illustrations on opposite page*].

One type of tube contains a single source of metal vapor near the positive-voltage end (the anode). The source is heated by a separate heater to about 250 degrees C. for cadmium and selenium to vaporize the metal into the discharge region. When the vapor reaches the discharge, some of the atoms are ionized and the positive metal ions are attracted toward the cathode by the electric field. The net result is a flow of metal vapor toward the cathode.

The discharge provides enough heat to keep the metal vapor from condensing in the bore region of the tube, but the metal does condense when it reaches the cathode region where there is no discharge and the temperature is lower. Thus there is a continuous flow of metal through the bore: the region where laser amplification is achieved. The electric current there provides the necessary population inversion. This flowing process is termed cataphoresis pumping. Its application to metal-vapor lasers was developed independently by the groups at Spectra-Physics and the RCA Laboratories and by Thomas Sosnowski at Bell Laboratories. In the helium-cadmium laser the process typically consumes cadmium at a rate of about one gram per 700 hours of laser operation.

The second type of system uses the metal over and over again. The vapor flows from the anode to the cathode in the laser discharge and then diffuses back to the anode through a return path. This laser was developed by Hernqvist at RCA to take advantage of special isotopes of cadmium that give a higher gain than the naturally occurring isotope distribution gives. The special isotopes are expensive, but only a small quantity is required in this tube since the metal is reused.

Another approach, which Leo Szeto and I developed at Bell Laboratories, employs a tube with a segmented bore. One or more cylindrical slugs of metal, with a hole drilled to match the bore region of the tube, are spaced between glass bore segments in the discharge region. The electrical discharge not only provides the high-energy electrons to excite the helium but also heats the metal to the proper vapor pressure. Cataphoresis then distributes the metal within the glass bore segments. The design, which is inexpensive, has been tested successfully with cadmium, selenium and zinc.

The fourth type of laser—the hollow-cathode laser—has a cylindrical metal cathode within which the metal is heated. Current flows from multiple anodes to the cathode and passes inside a slotted region of the cathode to produce a uniform glow in the cathode area. This kind of laser is more suited to the charge-transfer type of excitation than to the Penning type. It has not yet achieved the levels of power that can be produced by the other three systems. The first continuous output from a hollow-cathode structure was achieved independently by groups in Japan, the U.S.S.R. and the U.S.

There are many potential applications for a small, inexpensive laser operating in the blue or ultraviolet regions of the spectrum. For example, many molecules are highly sensitive to radiation in this region, and by changing the energy states of the molecules with laser light one can evaluate the chemical structure of the substance. Similarly, many photodetectors reach their maximum response in blue light, and so they could be used with blue-light lasers in such applications as scanning a page of text, translating the different intensities of black and white into electrical signals and transmitting them to a distant terminal, where the material could be reproduced.

One application under study at Bell Laboratories by Fritz Froehlich and his colleagues uses a small ultraviolet helium-cadmium laser similar to the segmented-bore design that I have described. The system, which is called a remote blackboard, would allow the transmission of handwritten information over a normal telephone line. The handwriting would be picked up from a blackboard or a piece of paper and transformed into an electrical signal that would be sent over the telephone line. At the other end the helium-cadmium laser would serve as the means of writing out the message on a small slide, from which the writing would be simultaneously projected on a screen. The laser is suited for this function because it can be focused to an extremely small spot on the ultraviolet-sensitive film in the slide. Galvanometers would deflect the beam to reproduce the handwriting.

One can also envision applications of the multicolored helium-selenium laser to devices designed for high-resolution color scanning and display. Moreover, recent developments in the field of dye lasers have indicated that the power from a small helium-selenium laser might be sufficient to pump a continuously operating dye laser. Such an application would lead to an inexpensive, tunable laser output in the red, orange and yellow regions of the visible spectrum.

Rotation in High-Energy Astrophysics

What is the source of the energy of pulsars, quasars and other strange objects? It may be gravitational energy converted into rotational energy as a large object contracts into a small one

by Franco Pacini and Martin J. Rees

The universe revealed by astronomy over the past three centuries has been a calm one, except for the occasional flare-up of a nova. This view has been shattered by the discovery of phenomena that involve rapid and violent changes. A typical example in our own galaxy is the behavior of the Crab Nebula in the constellation Taurus, the expanding debris from the explosion of a supernova described by Chinese astronomers in A.D. 1054. This object emits radiation at all wavelengths from radio waves to X rays, and its total output is equivalent to the power of 100,000 suns. The radiation is generated mainly by the synchrotron process, that is, by fast electrons spiraling outward along the lines of force in the nebula's magnetic field. Outside our galaxy there are more spectacular explosive phenomena, involving some radiation mechanism that is not connected with stars. Certain extragalactic objects even have features strikingly like those of the Crab Nebula; for example, the galaxy NGC 1275 has a filamentary structure. All these phenomena seem to have the same basic ingredients: fast particles and strong magnetic fields.

The central star of the Crab Nebula is a pulsar: an object that emits radiation in rapid pulses. It seems worthwhile to explore the notion that the Crab Nebula pulsar (and pulsars in general) may be the prototype for all the phenomena of high-energy astrophysics, enabling us to study at relatively close hand within our own galaxy the same mechanisms that operate on a grander scale in quasars and extragalactic radio sources.

A normal star depends on the energy released when hydrogen nuclei, or protons, are fused into heavier nuclei deep in the star's interior. There is an upper limit to the efficiency of nuclear fuels, set by the nature of the forces that hold the nuclear particles together. The maximum efficiency is .8 percent and corresponds to the maximum energy binding a particle within an atomic nucleus. Most of the energy generated in a star comes from hydrogen nuclei being converted into helium nuclei under conditions where each hydrogen nucleus has an energy of only a few thousand electron volts. It is difficult to see how such relatively modest energies could give rise to the fast electrons that generate synchrotron radiation, which travel at velocities close to the speed of light. Another source of energy would seem to be needed. That source could be gravity.

In stars such as the sun gravitational energy is rather unimportant. The gravitational energy that the sun would have released at its birth when it contracted from a diffuse cloud of gas to its present size would not have been sufficient to maintain its present luminosity for much more than 10 million years. The nuclear energy of the sun, on the other hand, can maintain our star in its present state for several billion years. In other words, the gravitational energy that binds an individual proton to the rest of the star is much less than the nuclear energy that binds the proton within a helium nucleus. Therefore the gravitational energy released by a normal star contributes only a relatively small amount to the overall energy budget.

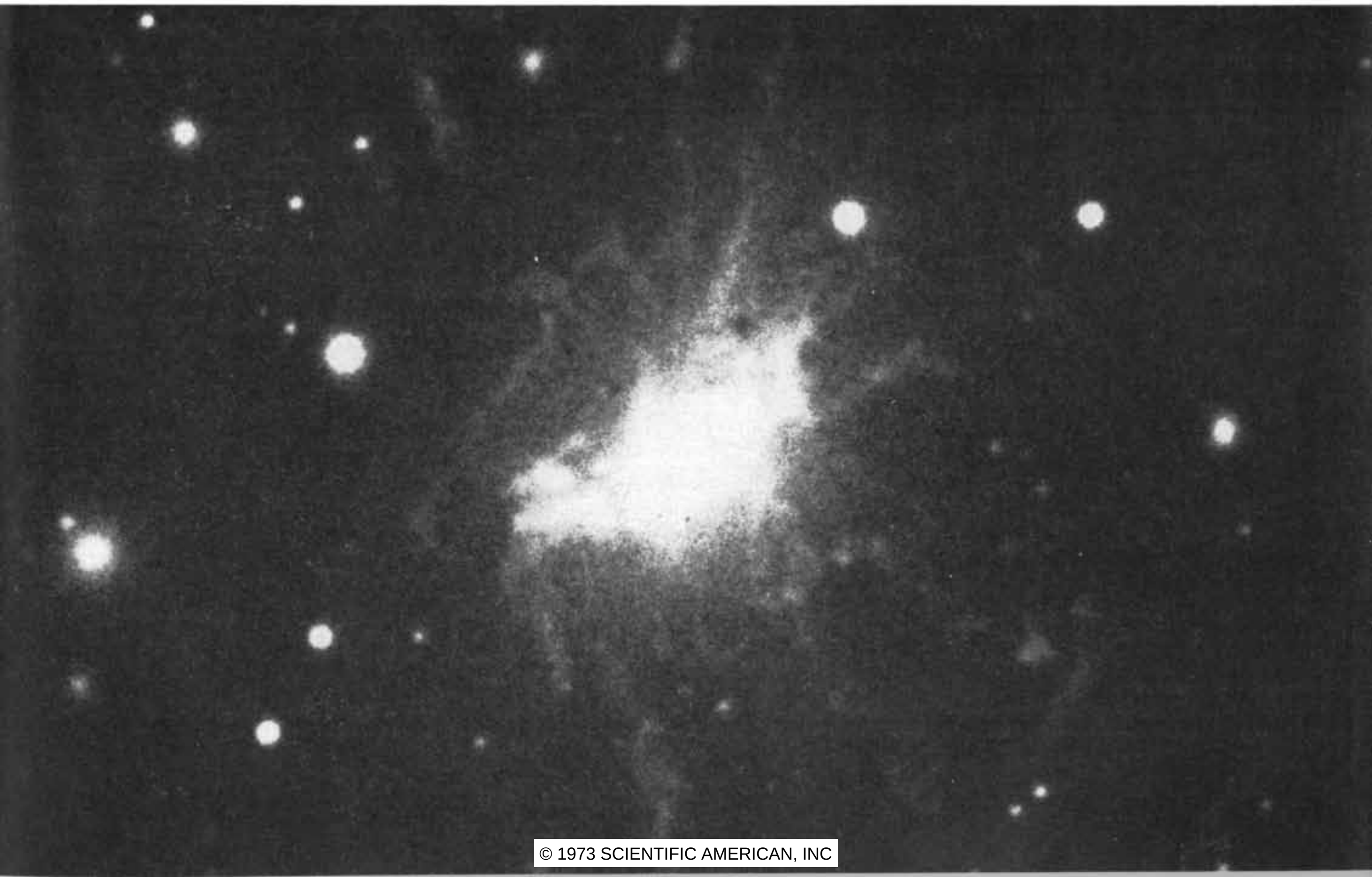
The gravitational binding energy of a star can be increased in two ways, however. First, it increases if more particles are added; the mass of the system grows, and gravitational force is directly proportional to mass. Second, it increases if the star shrinks; the attractive force between particles is inversely proportional to the distance separating them, and the binding energy depends on the inverse first power of this distance. Thus if the particles are compressed so that they are twice as close together, the star has twice the gravitational binding energy that it had before.

For a star with the mass of the sun the gravitational binding energy would exceed the nuclear binding energy if the star contracted to a radius of less than 100 kilometers. If the sun eventually becomes a neutron star with a radius of some 10 kilometers, then during its evolution it would have released more gravitational energy than it could conceivably have derived from nuclear reactions. As much as half of this energy may be available in the form of rotation. Therefore the "dead" star now begins a new life with more energy at its disposal than it ever had before! Gravitational energy can also overwhelm nuclear energy in supermassive stars: hypothetical objects more massive than a million suns that may play a role in the active nuclei of galaxies.

Two general features of gravitational collapse are of key importance in high-energy astrophysics. First, if the initial system is rotating, then as the gravitational field contracts it does work against the centrifugal forces. Angular momentum must be conserved; therefore the gravitational energy is gradually converted into the kinetic energy of rotation. Second, if there are internal electric currents generating a magnetic field, the energy associated with these currents increases as they are compressed; this energy must be supplied by gravitation. For example, the sun rotates about once per month and has a large-scale magnetic field of approximately one gauss at its surface (about the same as the earth's field). If the sun were squeezed to the size of a neutron star, it would spin at a rate of 1,000 rev-

CRAB NEBULA (*above*) in the constellation Taurus is the filamentary remnant of a supernova that exploded in our galaxy in A.D. 1054. Electrons spiraling at nearly the speed of light in the nebula's strong magnetic field emit synchrotron radiation at all wavelengths from radio region of the spectrum to the X-ray region.

GALAXY NGC 1275 (*below*) has a filamentary structure that strongly resembles the Crab Nebula, and it too emits synchrotron radiation from electrons spiraling through its magnetic field. NGC 1275 is one of a number of such objects outside the galaxy that behave very much like the Crab Nebula but on a vastly larger scale.

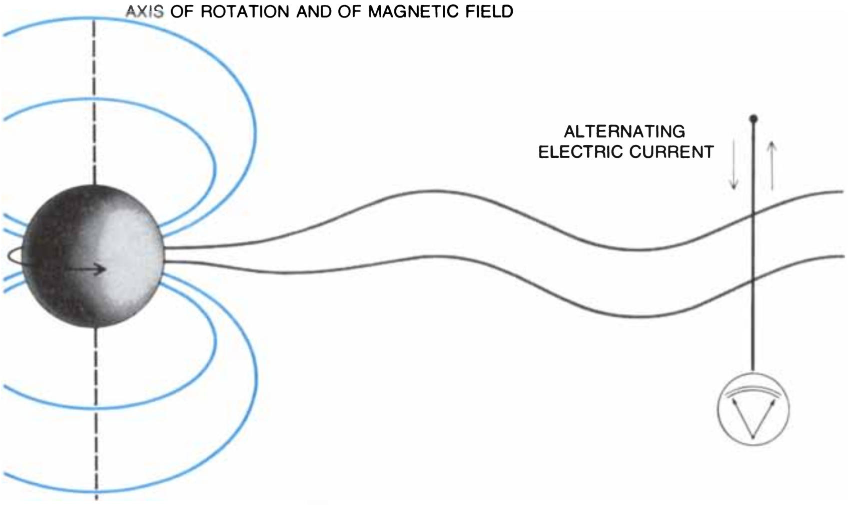


olutions per second and acquire a magnetic field of 10 billion (10^{10}) gauss.

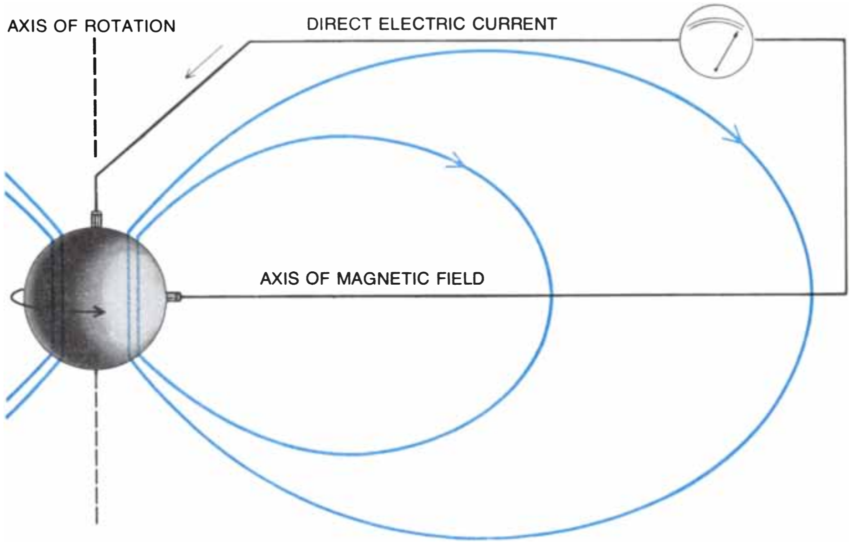
It is the combination of rapid rotation and enormous magnetic fields resulting from gravitational contraction that is now widely believed to play a basic role in the phenomena of high-energy astrophysics. The best evidence came with the discovery of pulsars and the suggestion that they are rapidly rotating neutron stars. This hypothesis, whose leading advocate was Thomas Gold of Cornell University, has been generally accepted. The most rapidly spinning pulsar, rotating 30 times per second, is the one in the center of the Crab Nebula. Its location is consistent with the theory that neutron stars form in the explosions of supernovas. The lifetime of the energetic electrons in the Crab Nebula responsible for the synchrotron radiation emitted at light and X-ray wavelengths is less than 1,000 years. Hence these particles cannot be the survivors of the initial explosion that created the nebula. In 1958 L. J. Woltjer, a Dutch astronomer whose doctoral thesis has become a classic text on the Crab Nebula, first emphasized that there must be some continuous source of fast particles inside the nebula. The Crab Nebula pulsar's greatest significance lies in the way it solves this long-standing problem of how the supply of energetic electrons is replenished. The range of masses and radii of neutron stars has been rather well defined by theoretical work. If it is known how fast a star is spinning and how quickly it is slowing down, it is relatively easy to compute how fast it is losing energy. The period of the Crab Nebula pulsar is increasing by about 15 microseconds per year, implying that the pulsar is losing energy 100,000 times faster than the sun. This rate agrees remarkably well with the energy required to maintain the synchrotron emission, implying that there must be some mechanism converting the rotational energy of the pulsar into fast particles with an efficiency that is certainly higher than the best value of about 1 percent achieved by particle accelerators in the laboratory.

What process provides the torque that can extract energy from a spinning neutron star? There is little doubt that the magnetic field plays a key role. A magnet spinning in a vacuum will radiate electromagnetic waves with the same frequency as its frequency of rotation, provided that the axis of the magnetic field does not precisely coincide with the axis of rotation. The waves carry energy away at a rate that depends on the strength of the magnetic field and the rate of rotation. If the Crab Nebula pulsar has an oblique magnetic field, that is, a field whose axis does not coincide with the axis of rotation, the strength of the field at the surface of the star would have to be about 10^{12} gauss to produce the observed rate at which the pulsar is slowing down [*see upper illustration at left*].

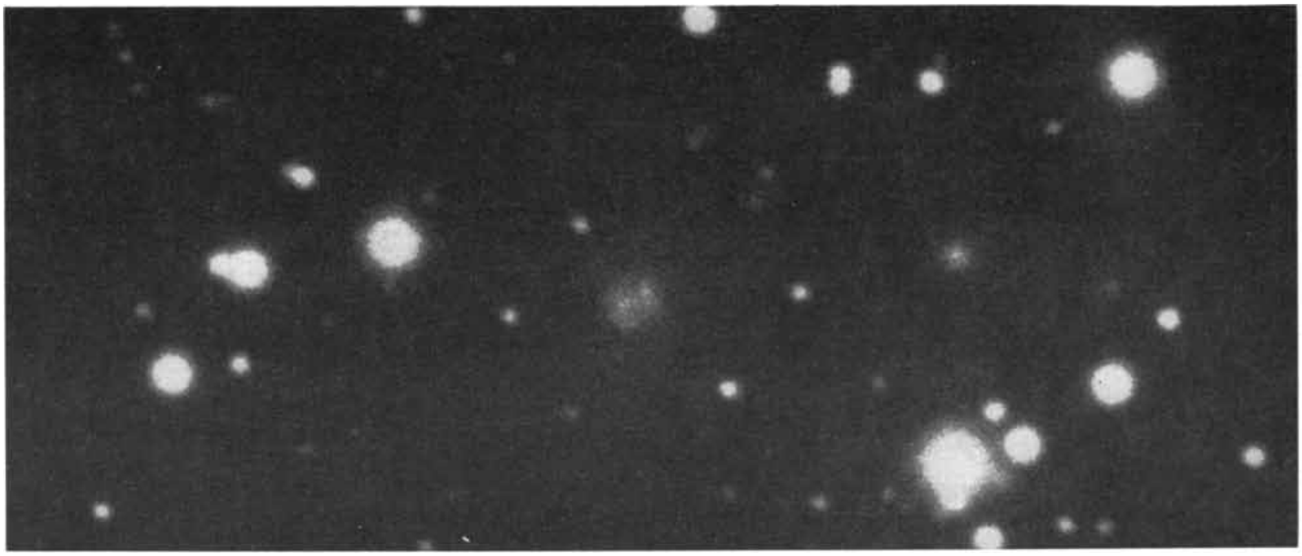
Although such a field is phenomenally strong by laboratory standards, it is well within the range that would be expected if the magnetic field in a normal star is "frozen in" and amplified when the star is compressed into a neutron



OBLIQUE DIPOLE FIELD is one mechanism that can extract energy from a spinning neutron star or a pulsar and cause it to gradually slow down. The axis of the star's magnetic field (*color*) does not coincide with its axis of rotation but is at some oblique angle to it. As the magnetized star spins, it radiates electromagnetic waves (*gray*) that have the same frequency as its frequency of rotation, provided that the star is in a good vacuum and not embedded in too much ionized gas. The waves accelerate electrons to velocities near the speed of light and cause them to emit synchrotron radiation. The amount of energy that is carried away depends on the strength of the star's magnetic field and on its rate of rotation.



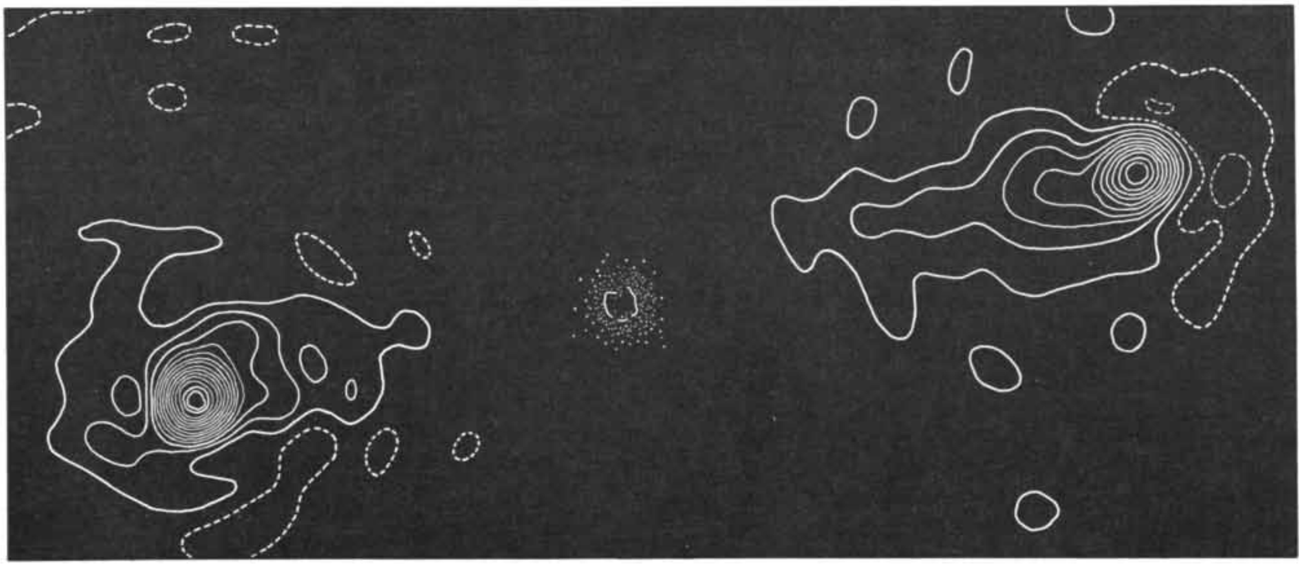
UNIPOLAR INDUCTOR is another mechanism that can extract energy from a spinning neutron star or pulsar. It will work even if the star is surrounded by too much plasma (ionized gas) for pure electromagnetic waves to propagate. If in the laboratory one spins a magnetized conducting sphere, and attaches one terminal of an external circuit to a point on the axis and the other to a fixed brush in contact with the sphere's equator, then a galvanometer connected into the circuit will register the passage of current and the spinning sphere will slow down. In the case of a neutron star the surrounding plasma plays the role of the wire of the external circuit and currents are again set up. The energy for this mechanism again comes from the rotational energy of the star, and the braking torque imparted to the star is about the same as it would be if electromagnetic waves carried away the energy. The main difference is that the unipolar inductor, unlike the oblique dipole-field mechanism, can also act for a star whose magnetic poles point along its axis of rotation.



RADIO SOURCE CYGNUS A appears as a faint object visible in the center of this photograph taken with the 200-inch reflecting telescope on Palomar Mountain. Originally it was speculated that its strong synchrotron radiation might be emitted as the result of two galaxies colliding. It is now believed that the radiation originates with some kind of activity intrinsic to an individual galaxy.

star. Detailed calculations by James E. Gunn and Jeremiah P. Ostriker of Princeton University and others have shown that very-low-frequency electromagnetic waves emitted by a spinning pulsar could be exceedingly efficient at accelerating charged particles. Any particle injected is immediately accelerated and would ride outward on the crest of the wave, as it were, at almost the speed of light [see "The Nature of Pulsars," by Jeremiah P. Ostriker; SCIENTIFIC AMERICAN, January, 1971]. It is still a controversial question whether or not the environment of the Crab Nebula pulsar is sufficiently close to being a vacuum to allow the presence of these low-frequency electromagnetic waves.

Even if the star is surrounded by too much plasma, or ionized gas, for pure electromagnetic waves to propagate, it can still lose energy by the cosmic equivalent of a unipolar inductor, a device first investigated by Michael Faraday. In the laboratory a unipolar inductor is a magnetized metal sphere. One terminal of an external circuit is attached to one of the sphere's poles of rotation and the other terminal is a stationary brush in contact with the sphere's equator. When the sphere is spun, a galvanometer in the circuit registers the passage of an electric current. In the case of a neutron star the surrounding plasma plays the role of the wire of the external circuit and electric currents are set up. The energy that drives this mechanism comes, as in the laboratory case, from the rotational energy of the spinning body. For a magnetic field of given strength the torque imparted by the unipolar inductor is about the same as it would be if the energy were carried away by electromagnetic waves. Thus the magnetic field of the star would still have to be about 10^{12} gauss. The single difference is that the unipolar inductor can operate in a system that is sym-



MAP OF CYGNUS A at the radio wavelength of six centimeters reveals that the source emits its radiation from two areas that are placed fairly symmetrically on each side of the central object. The contour lines of the map represent gradations in the surface brightness of the radio source. This map of the source was compiled by Simon Mitton and Martin Ryle of the University of Cambridge.

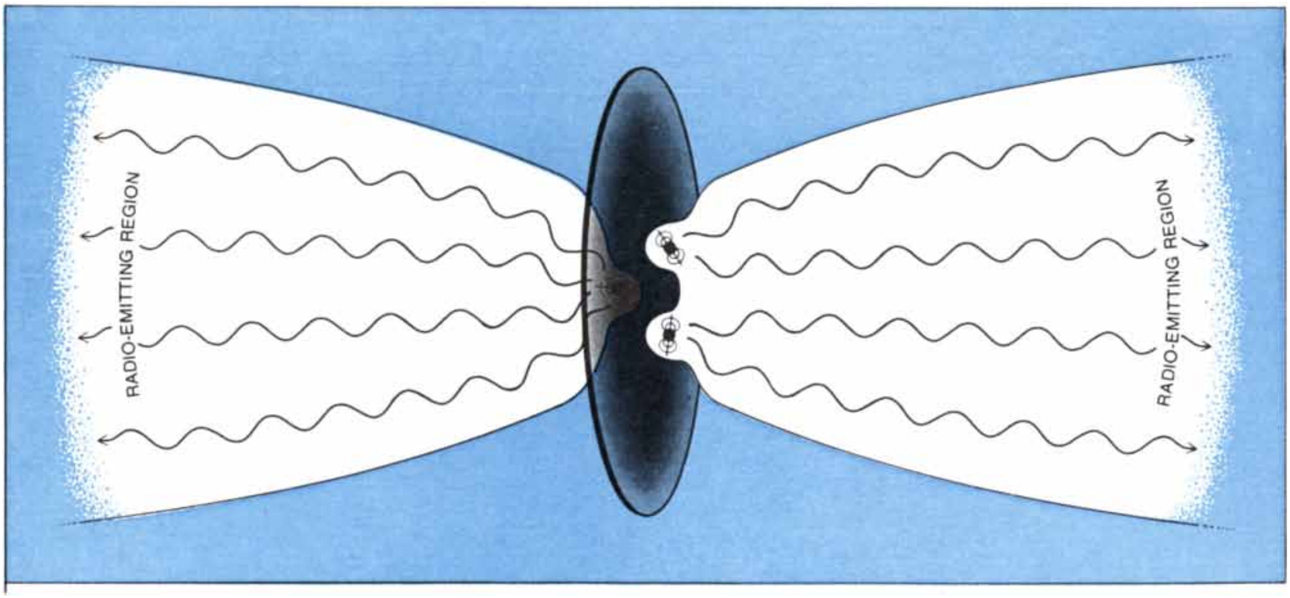
metrical around its axis of rotation. The Crab Nebula pulsar could generate potential differences as large as 10^{14} volts by the unipolar-inductor mechanism, and particles would attain high energies by a process that is essentially electrostatic, as in a Van de Graaff generator. The forces exerted by the electric fields at the surface of the neutron star would exceed the force of gravity by a factor of millions or even billions. As Peter M. Goldreich and William H. Julian of the California Institute of Technology have shown, this means that the star itself can supply enough electrically charged particles to maintain a high conductivity in its environment.

How will a system such as the Crab Nebula evolve in the future? After some 10,000 years the supernova remnant will have expanded until it has merged into the general interstellar medium, and the high-energy particles produced by the pulsar will contribute to the general background of cosmic rays in the galaxy. Although the rotation of the pulsar will slow down, the pulses may continue for several million years. The 60 or 70 other known pulsars were presumably formed in events similar to the supernova of A.D. 1054. Little trace of their filamentary debris survives, except for the nebulosity surrounding the pulsar in the constellation Vela. The Vela pulsar has the second-shortest period known, and it may therefore be the second-youngest.

Eventually the electromagnetic output of a pulsar would become so small that the radiation pressure associated with it could no longer hold the interstellar medium at bay. The pulsed radio emission would then be quenched. The neutron star would cease radiating unless it were heated by the interstellar matter falling into it, which would cause it to emit in the X-ray region of the spectrum. If it were isolated in the interstellar medium, the rate at which it accreted matter would be very low and its X-ray luminosity would be commensurately so. Such a faint object would be detectable only out to about 10 parsecs (33 light-years) from the sun. The collective contribution from all the dead pulsars in the galaxy (which could number as many as a billion) could, however, produce a significant X-ray background. On the other hand, if a dead pulsar were located in a dense cloud of interstellar matter or were a member of a double-star system, then the rate at which matter accreted would be greater and the X-ray emission would be enhanced. It is believed that many of the discrete X-ray sources in the galaxy, which appear to be compact objects in a close, almost grazing, orbit around a companion, can be explained along these lines. The X-ray emission results mainly from the impact of gas, accelerated by gravity to about half the speed of light, on the crust of the neutron star. Slow rotation might still induce periodic variations in the observed X-ray luminosities if, for example, the magnetic field could channel accreted matter onto certain parts of the star's surface. Although the spin rate may still provide the "clock," it is the gravitational energy from the falling material, rather than the rotation, that probably powers the X-ray sources.

There have been several attempts to adapt similar ideas to account for the violent activity that has been detected in recent years in the nuclei of galaxies. The luminosity of any one galactic nucleus can range up to 10^{13} times the luminosity of the sun at visible wavelengths. The total power emitted is perhaps even more if the nucleus radiates strongly in the infrared, which is mostly barred to observers on the earth by water vapor in the earth's atmosphere. Such luminosities are some 100 times greater than the luminosity of a galaxy such as ours. Moreover, this colossal output of power is concentrated in a small central region. It is observed to vary, at both light and radio wavelengths, over a period of months or even less. If the size of the volume emitting the radiation exceeded a few light-months, these variations would be smeared out, since light takes a finite time to travel from one side of the source to the other. Therefore it looks as though in these galaxies a region not much larger than the solar system can release 100 times as much power as the whole of a normal galaxy.

One example of this type of object is



MANY PULSARS created by frequent supernova explosions in the nucleus of a galaxy could offer one explanation for strong extragalactic radio sources. Each newly formed pulsar might create in the surrounding gas an expanding cavity (*white*) that contains only electromagnetic fields and fast electrons. If the gas in the central part of the galaxy (*gray*) is rotating, it will tend to flatten into a disk. Electromagnetic waves (*black*) will escape from the galactic nucleus preferentially along the axis of rotation. These waves, focused into two well-collimated beams, would eventually interact with the intergalactic medium (*color*) and generate fast electrons radiating at radio frequencies. One can perhaps account for symmetry and other properties of double radio sources along these lines.

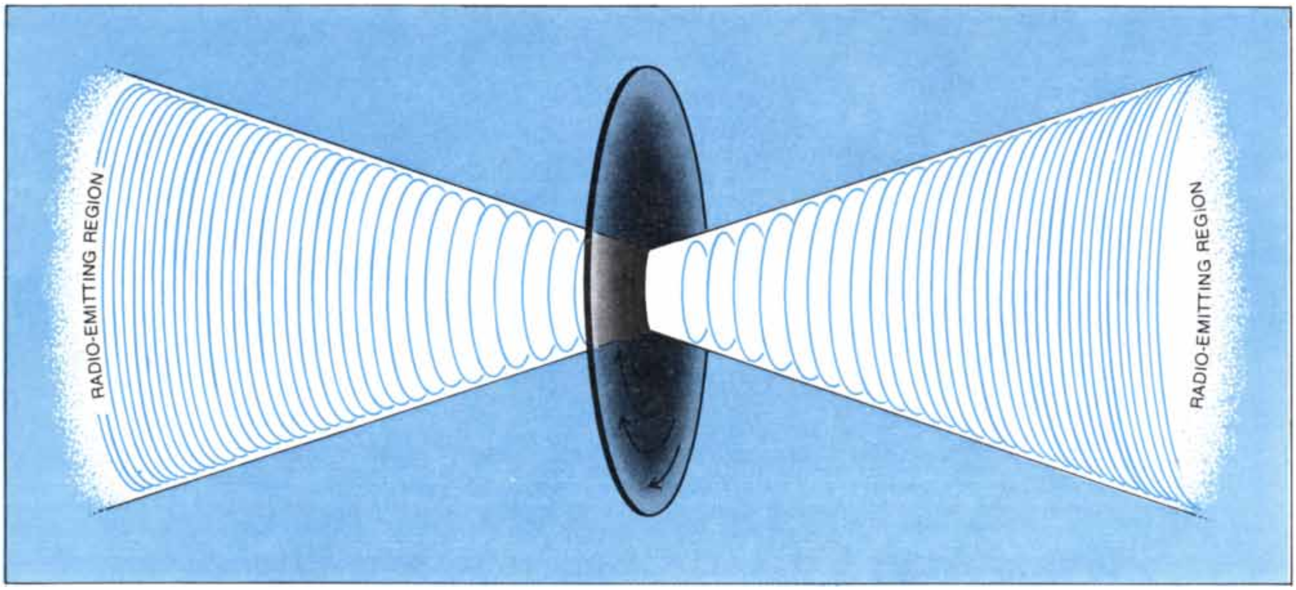
the Seyfert galaxies: spiral galaxies with very bright nuclei where the light comes mainly from glowing gas and fast particles rather than from stars. A second example is the quasars, which most astronomers believe are "hyperactive" galactic nuclei, manifesting far more powerful activity than Seyfert nuclei, whose luminosity swamps the contribution from stars in the surrounding galaxy. Quasars are often associated with strong, extended radio sources, which characteristically radiate from two major areas placed fairly symmetrically on each side of a central object. It is assumed that the electrons emitting the synchrotron radiation at radio wavelengths somehow originate in the parent object. These extended double radio sources are also found in connection with elliptical galaxies, whose nuclei are not conspicuously active in other ways. The strongest radio sources, such as the source in the constellation Cygnus designated Cygnus A, were the first instances of violent activity outside the galaxy to be discovered. It was originally theorized that they might represent collisions between galaxies, but it is now clear that they involve some kind of activity intrinsic to an individual galaxy.

The nature of this activity in radio galaxies, quasars and related objects is still mysterious; many people believe it may involve some fundamentally new principles of physics. This raises the exciting prospect that astronomical studies might reveal some new laws of nature that would elude discovery in laboratory experiments. It would be rash to accept this conclusion, however, without first exploring all the possibilities encompassed by conventional physical principles. The best hope of interpreting these phenomena along conventional lines lies perhaps in applying the same ideas that seem to have worked successfully for the Crab Nebula.

One possible explanation for the strong extragalactic radio sources is that the frequency of supernova explosions is much higher in active galactic nuclei than in ordinary galactic nuclei (perhaps because the constituent stars are different) and that as a result energetic young pulsars are being produced continuously. The Crab Nebula shows that a pulsar can create a near-vacuum in its surroundings by exerting an overwhelming pressure of electromagnetic radiation. Something similar could happen in the gas enveloping the pulsars in the nuclei of galaxies: each newly formed pulsar might create an expanding cavity that contains only electromagnetic fields and fast electrons. Detailed studies of the nature of the variation at radio wavelengths suggest that such clouds of particles and fields expand and fade over an interval of a few months. Thus the cavities can perhaps be identified with the successive radio outbursts. The outbursts can also create and maintain the reservoir of particles needed to account for the luminosity of the source at other wavelengths.

If the ambient matter in the galactic nucleus is spinning, it will tend to flatten out into a disk. The fast particles and magnetic fields will then escape from the nucleus preferentially along the axis of rotation. That may be part of the explanation for the directional symmetry of the extended double radio sources. A few astrophysicists even believe that low-frequency electromagnetic waves generated by spinning pulsar-like bodies in the nucleus may be channeled along tubes in the intergalactic medium—rather as high-frequency waves are channeled in a man-made waveguide—to accelerate the electrons in the radio-emitting regions.

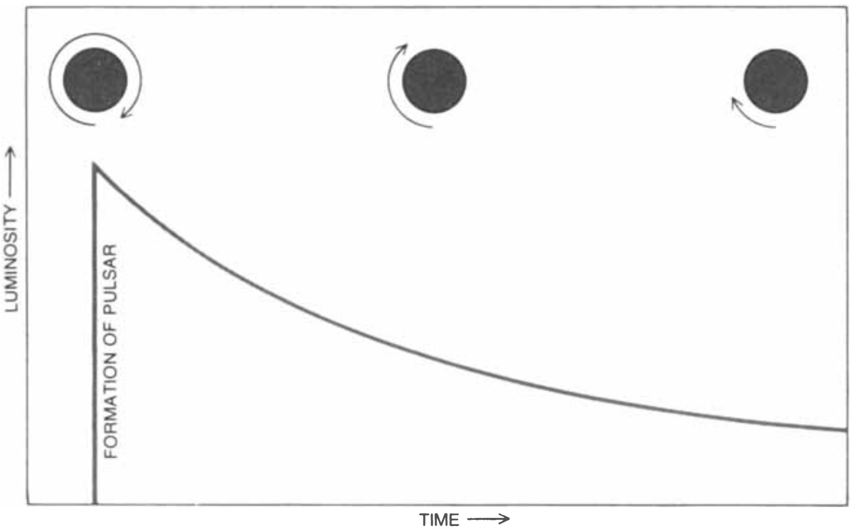
Instead of regarding the galactic nucleus as being composed of many small objects each with the mass of an ordinary star, one can consider the alternative possibility that all the mass may reside in one or a few supermassive stars, each with 10^8 (or more) times the mass of the sun. The gravitational collapse of supermassive stars was first suggested by William A. Fowler of Cal Tech and Fred Hoyle of the University of Cambridge to explain the energetics of strong radio sources. Such an object would only vaguely resemble an ordinary star. There would be a delicate balance between its gravity and its internal radiation pressure. It would be unstable unless it pos-



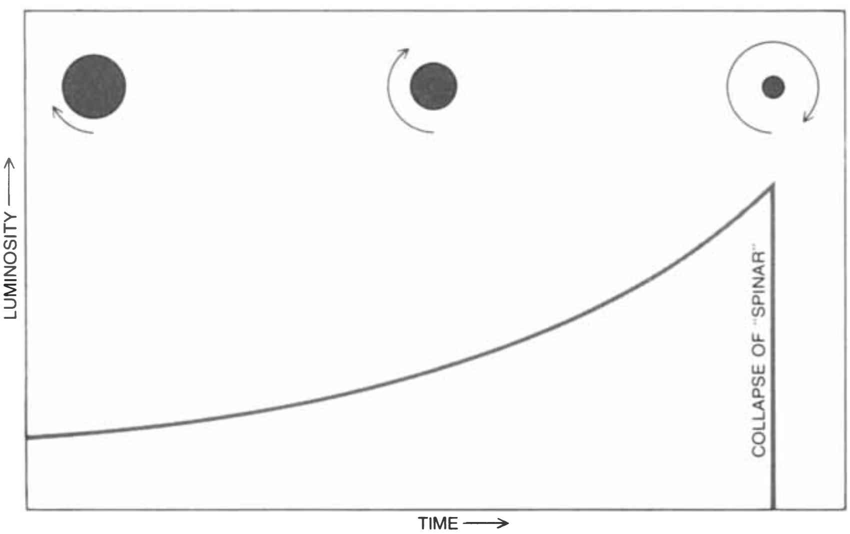
ENORMOUS BODY in the middle of a galaxy, with a mass 100 million times the mass of the sun, could offer an explanation for extended double radio sources alternative to the many-pulsar model. Such a supermassive star would lack the rigidity of a neutron star and would not rotate as a solid body; its inner parts would tend to rotate faster than its outer parts. The lines of force of its magnetic field (*lines in color*) would be coiled up and amplified by this differential rotation. The field might explode outward like a spring on both sides of the disk. Strong synchrotron radiation at radio wavelengths would be emitted by fast electrons in the regions where the field is compressed and strengthened by interaction with intergalactic medium (*light surrounding color*).

sessed either a magnetic field or rotation, but of course it is a rotating magnetized supermassive star that would be indicated if there were indeed any similarity to processes occurring in the Crab Nebula.

Philip Morrison of the Massachusetts Institute of Technology was the first to suggest that the almost regular variation of the light from the quasar 3C 345 might establish a direct analogy between quasars and pulsars, with the difference that in quasars the time scale of the variation is several months rather than of the order of a second. If a supermassive object had a magnetic field at its surface of 10^5 to 10^6 gauss, magnetic stresses would carry away rotational energy at a rate sufficient to power a quasar. Again variable electromagnetic fields induced by the rotation would accelerate particles to velocities near the speed of light. These particles would be responsible for the copious emission in the various wavelengths. Such a rotating supermassive object might be called a "spinar." There would be one important difference between a pulsar and a spinar. In a pulsar rotation does not contribute appreciably to the equilibrium of the object, but in the supermassive spinar centrifugal forces provide the main support against the entire object's collapsing. When rotational energy is extracted from a supermassive star, it contracts. As a result the object spins faster and increases its rotational energy. Thus whereas a pulsar fades in its old age, a spinar gets brighter. One complication arises from the likelihood that a supermassive star, which lacks the rigidity of a neutron star, will not rotate as a solid body. If different parts of the object rotate at different speeds, the lines of force of the associated magnetic field will be coiled up and amplified. Several workers have proposed that such a field may explode outward like a spring, and that this process could be the explanation for double radio sources [*see illustration on preceding page*].



PULSAR SPINS SLOWER and becomes dimmer as it ages. After the original star explodes violently as a supernova, the remnant becomes a neutron star and a pulsar. It is supported against gravity mainly by nuclear forces, and rotation does not contribute appreciably to its equilibrium. The youngest pulsars spin fastest and radiate most strongly. Gradually they fade and slow down as electromagnetic torques in their medium carry away their energy.



"SPINAR" SPINS FASTER and becomes more luminous as it ages. Centrifugal force from rotation rather than internal pressures would support this hypothetical supermassive object against collapse. When rotational energy was extracted, the spinar would contract; as a result it would spin faster and increase its rotational energy. When a substantial fraction of its mass had been radiated away as energy, it would contract to the point where the escape velocity at its surface was close to velocity of light. It would then collapse into a black hole.

An object largely supported by rotation would inevitably be highly flattened by centrifugal forces. For this reason the evolution and stability of rotating disks have recently received wide attention from theorists. Disks are subject to various instabilities; a supermassive disk could fragment into a swarm of smaller disks or pulsar-like objects bound to one another gravitationally but all radiating independently. Supermassive stars, disks and pulsars are in any case rather similar in electromagnetic behavior. In this sense the two proposed models of the nuclei of galaxies—a single supermassive object or many pulsars—may represent different stages in the evolution of the same kind of system.

As we have seen, a spinning supermassive star becomes progressively more luminous as it contracts. Clearly this process cannot continue indefinitely. Eventually a substantial fraction of the spinar's mass will be radiated away as energy. The object will contract so much that its size will approach what is known as the Schwarzschild radius: the radius at which the gravitational attraction of the object is so strong that the escape velocity of particles is equivalent to the speed of light. At this point the effects of the general theory of relativity become crucial. The internal pressure of any object with a mass substantially exceeding the mass of the sun cannot withstand gravity after the available sources of heat are exhausted. Thus it would seem that after a supermassive object had lost most of its angular momentum it would eventually collapse to form what is now called a black hole. It is also conceivable that a swarm of pulsars could eventually coalesce and suffer the same ultimate fate.

These considerations suggest that very massive black holes exist in the center of many galaxies. Indeed, since the duration of violent activity is usually thought to be from a hundredth to a ten-thousandth the age of a galaxy, it is possible that all galactic nuclei could at some time have passed through a stage of violent activity, and that there may be a massive black hole in the center of even a quiescent galaxy such as ours. Such black holes would accrete matter from the surrounding gas. If the gas has angular momentum, the accretion cannot proceed on straight lines toward the center of the black hole. The material will tend to orbit the central body and make up a thin disk. The outer parts of the disk will spin slower than the inner parts. Viscous friction will therefore tend to brake the inner parts and transport angular momentum outward. As gas in the inner part of the disk loses angular momentum, it swirls inward closer to the center. In this way material could gradually spiral inward at a rate controlled by the strength of the viscous forces until it was eventually swallowed by the black hole. The viscosity would dissipate energy into heat, magnetic fields or fast particles, and so the disk would radiate. The energy, which comes from the gravitational energy released as matter falls inward, can amount to as much as 40 percent of the total rest-mass energy of the accreted material. Attempts have been made to interpret the nonstellar radiation from some galactic nuclei in terms of this kind of disk. The accretion of matter by black holes is the most efficient mechanism yet proposed for converting rest mass into radiation, apart from such exotic processes as the total annihilation of matter when it comes into contact with antimatter.

All the processes we have described intimately involve rotation. In all cases it seems to be rotation that determines how fast a system can utilize its gravitational energy, and the influence of rotation tends to be amplified as a system contracts. In view of the obvious role of rotation in determining the features and evolution of such spectacular large-scale systems as spiral galaxies, it would be surprising if rotation failed to be dominant in compact systems as well. Although some of the detailed processes on which people now speculate may turn out to be irrelevant, these general ideas seem to offer our best hope of understanding the phenomena of high-energy astrophysics within the framework of conventional physics.

MATHEMATICAL GAMES

Up-and-down elevator games and Piet Hein's mechanical puzzles

by Martin Gardner

Elevators, unlike cars, trains, planes, ships and other common modes of transportation, have been unduly neglected by recreational mathematicians. This month we undertake to rectify the situation by considering four unusual elevator problems. The first three were provided by Donald E. Knuth, a computer scientist at Stanford University and author of a classic seven-volume work in progress titled *The Art of Computer Programming.* Before discussing two combinatorial problems that appear for the first time in his third volume (published in January by Addison-Wesley), we consider a well-known probability paradox with a startling generalization that Knuth discovered a few years ago.

George Gamow and Marvin Stern introduced the elevator paradox in the prologue to their little book *Puzzle-Math* (Viking, 1958). Gamow once had an office on the second floor of a seven-story building in San Diego and Stern had an office on the sixth floor. When Gamow wanted to go up to see Stern, he noticed that in about five cases out of six the first elevator to stop on his floor was going down. It seemed as if elevators were being manufactured on the roof and then sent down the shafts to be stored in the basement. For Stern the situation was the opposite. When he wanted to go down to see Gamow, about five times out of six the first elevator to arrive was on its way up. Were elevators being fabricated in the basement and then sent to the roof to be carried off by helicopters?

The explanation, as Knuth pointed out later, requires a few idealizing assumptions. Suppose each elevator travels independently in continuous cycles from bottom floor to top and back again, moving with constant speed and with the same average waiting time on each floor. Thus at the time a button is pushed on any floor we can assume that each elevator is at a random point in its cycle.

For a single elevator, calculating the probability that it is on its way down when it stops on a given floor is quite easy. Stern, on the sixth floor, has five floors below and one above; therefore the probability is 5/6 that the elevator is below him and will be moving up when it arrives. Gamow, on the second floor, has five floors above and one below; therefore the probability is 5/6 that the elevator is above him and will reach his floor on its way down. Gamow and Stern explained this in their book, but then they made a slip. If there is more than one elevator, they wrote, the probabilities "of course remain the same." The slip is understandable because the statement seems so intuitively true. Apparently Knuth was the first to realize that it is not true at all! Indeed, as the number of elevators approaches infinity the probability that the first elevator to stop on any floor (except the top or bottom floors) is going up (or down) approaches exactly 1/2—a rather unexpected result. Yet the probability (for, say, the second floor) remains 5/6 for every individual elevator, and all elevators are equally likely to be the next to arrive.

The solution for two or more elevators is complicated by conditional probabilities. As Knuth puts it: "The choice of which elevator is first to arrive on the second floor is partly contingent on whether it was above us or below, since an elevator that is below the second floor when we begin to wait is likely to arrive ahead of an elevator that is above (all other things being equal)." In his 1969 paper [*see "Bibliography," page 124*] Knuth analyzes Gamow's situation as follows: Consider the portion of an elevator's route that starts at the fourth floor, then goes down to the first floor and up to the second, a total of 4/12 = 1/3 of the entire route. During the first half of this portion the elevator stops next at the second floor going down, and during the other half it will next stop going up. Therefore we may call it the unbiased portion, since it is not biased toward up or down.

If there are n elevators, Knuth now distinguishes two cases:

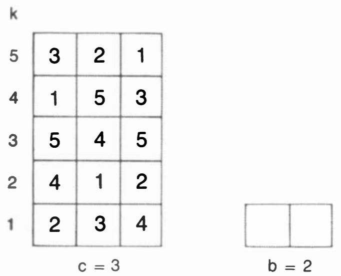
1. No elevator is in the unbiased portion. The probability of this is $(2/3)^n$, since it is 2/3 for each elevator. The next elevator to stop on the second floor will be going down.

2. At least one elevator is in the unbiased portion. The probability is $1 - (2/3)^n$. We can ignore any elevator outside the unbiased portion, since one of those in the unbiased portion will necessarily reach the second floor first. In this case the elevator will be going down with probability 1/2.

Combining these results gives a probability of $(2/3)^n + \frac{1}{2}(1 - (2/3)^n) = \frac{1}{2} + \frac{1}{2}(2/3)^n$ that the first elevator to arrive on the second floor will be going down. If there are just two elevators running in Gamow's seven-story building, the first elevator to stop at the second floor will be headed downward with probability $\frac{1}{2} + 2/9 = 13/18$. This is slightly less than 5/6, so that Gamow's chances of catching an up elevator have improved. If there are seven elevators, the probability of an elevator's going down would be 2,315/4,374, which is not far from 1/2.

Knuth gives the general formula for any building by defining p as the distance from a given floor to the bottom divided by the distance between top and bottom floors. For Gamow p is 1/6, for Stern p is 5/6. The general formula for all values of p between 0 and 1 is

$$\tfrac{1}{2} + \tfrac{1}{2}(1 - 2p)\,|\,1 - 2p\,|^{\,n-1}.$$



u_k/b	d_k/b	$[u_k/b]$	$[d_k/b]$
0	3/2	0	2
3/2	5/2	2	3
5/2	3/2	3	2
3/2	3/2	2	2
3/2	0	2	0
		9 +	9 = 18

Five-story-building elevator problem

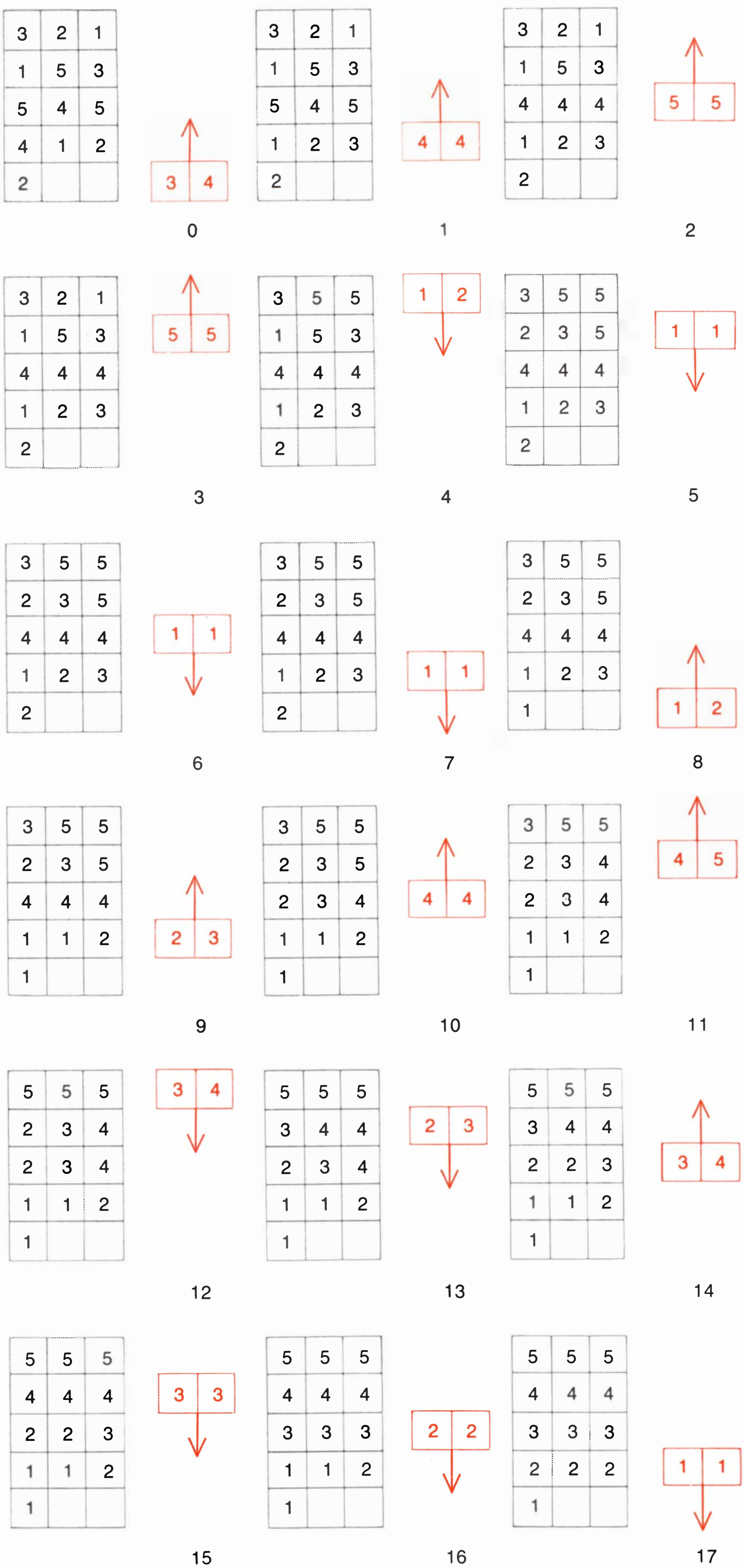
The pair of vertical lines indicate the absolute value of the expression between them. The probability approaches 1/2 as n, the number of elevators, approaches infinity.

Our second elevator problem is from the third volume of Knuth's series, a book that deals entirely with computer techniques of sorting and searching for information. Like its two predecessors, it is comprehensive in scope, written in a clear, informal style (although at times it is necessarily terse and technical) and rich in humor, historical data and problems of great recreational interest. On pages 11 through 72, for instance, Knuth brilliantly summarizes almost everything known about the combinatorial properties of permutations, a topic that ties in with scores of classic puzzle problems. The book's exercises concern such entertaining topics as solitaire card games, shuffling, anagrams, snowplows, the design of tennis tournaments (including Lewis Carroll's flawed efforts to find a design that does the best possible justice to the second-best player), rook problems, sorting puzzles, the unsolved weight-ranking problem, the Josephus problem, parking problems, Fibonacci numbers, the odd "tableaux" of Alfred Young (which have a curious relevance to the eightfold way of particle theory) and a hundred other things that lead straight into recreational mathematics.

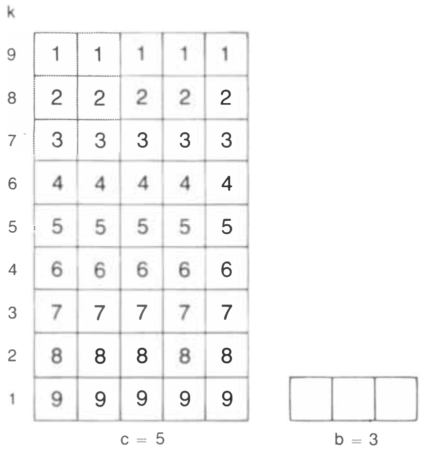
Here we are concerned with pages 357 through 360, where Knuth regards the elevator as a model of one-tape computer sorting. A building has n floors, each holding exactly c people. There is a single elevator that carries at most b people. We assume that the building is full (contains cn people). Exactly c persons want to go to each floor: c to the first floor, c to the second floor and so on. Some people may already be on their desired floor, but it is more interesting to assume that all or most are misfits who want to be on another floor.

The elevator always starts at the bottom. It moves up and down, loading and unloading passengers, until each person is where he wants to be. The elevator then returns to the first floor. A movement of the elevator from any floor to the next floor above or below will be called a unit trip. The problem is to find an algorithm that will sort all the people in a minimum number of unit trips. This operation is equivalent, of course, to minimizing the distance traveled or (assuming a constant elevator speed) to minimizing the time required for the sorting.

As Knuth points out, the people correspond to records that are to be com-



Richard M. Karp's algorithm



Nine-story-building elevator problem

puter-sorted. The building is the tape, the floors are blocks on the tape and the elevator is the computer memory. A computer can do such things as duplicate records or chop them into parts to be stored temporarily in different blocks. It turns out, however, that a clever algorithm discovered by Richard M. Karp enables the elevator to do its job with peak efficiency without having to duplicate or partition any passenger.

Let k be the number of the floor, u_k the number of misfits on k and all lower floors who want to go higher than k, and d_k the number of misfits on k and all higher floors who want to go lower than k. It is not hard to see that $u_k = d_{k+1}$. For example, suppose k equals 3. The theorem states that all people on floors 3, 2 and 1 who want to go above Floor 3 are equal in number to those on Floor 4 and higher who want to go below Floor 3. (It is like the old wine-and-water problem. In a filled building the people who go up from a bottom portion of the building must be replaced by the same number of people in the top portion who want to go down.) Both u_n (misfits on the top floor) and d_1 (misfits on the first floor) are, of course, zero, since no one wants to go above the top floor or below the first floor.

Because the elevator holds at most b people it must make at least $\lceil u_k/b \rceil$ trips from Floor k to the next floor above, where $\lceil\ \rceil$ symbolizes the roundup function (the value is rounded up to the nearest integer). Similarly, the elevator must make at least $\lceil d_k/b \rceil$ trips from k down to the next floor below. If we now calculate $\lceil u_k/b \rceil$ and $\lceil d_k/b \rceil$ for each floor, the sum of all these integers will be the least number of trips that the elevator must make to sort everyone.

Karp's algorithm achieves this minimum if u_k is not zero for any floor except the top one and provided that the number of people each floor can hold is not less than the number the elevator can hold. The procedure calls for the assumption that the elevator is always in either the UP state or the DOWN state. It starts in the UP state and repeats the following algorithm until everyone is sorted:

1. When the elevator is in the UP state, if anyone (in the elevator or on the floor where it has just stopped) wants to go up, fill the elevator with those of the highest destination, with all others remaining on the floor, then move the elevator up one floor. Otherwise change to the DOWN state.

2. When the elevator is in the DOWN state, fill it with those people of the lowest destination (who are on the elevator or on the current floor) and move the elevator down one floor. Then change the elevator to the UP state if there are no misfits on lower floors who want to go to the new current floor or higher.

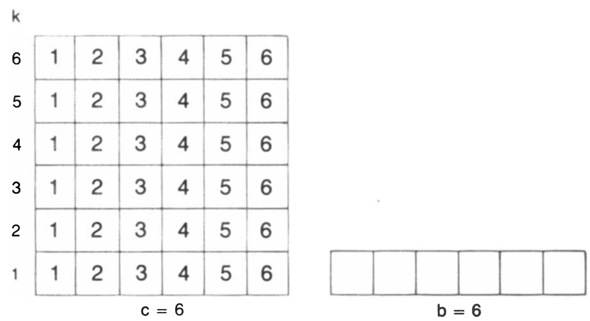
To see exactly how this operates, consider a five-floor problem [*see illustration on page 106*]. Each floor holds three people. Each person is represented by a numeral that indicates the floor he wishes to go to. The empty elevator on the right can hold only two people. In this problem all the people in the building are misfits except for one 2-person who is already on Floor 2. In order to calculate the minimum distance the elevator must travel first list the u_k/b and d_k/b values for each floor and then list these values rounded up [*see illustration*]. Note the positions of the zeros and the fact that the sequence of values for u_k/b is repeated in the d_k/b column, except that the sequence starts one floor higher. This repetition is true of all such charts and is a consequence of the theorem $u_k = d_{k+1}$. The sum of the rounded-up values is 18, so that we know the elevator must make at least 18 unit trips to accomplish the sorting and then return to the first floor.

The illustration on the preceding page shows what happens when we apply Karp's algorithm. (The final step is not shown.) Observe that occasionally people are taken off the floor on which they wish to stay. In some cases the procedure will carry a person in one direction when he wants to go in the other. "This represents," Knuth writes, "their sacrifice to the common good."

To get a feeling for the spooky way Karp's algorithm does its job, readers are urged to work out the problem of sorting 45 people in a nine-floor building with an elevator that holds three people [*see top illustration on this page*]. First calculate the minimum number of unit steps needed. Then draw the building on a sheet of cardboard, fill the rooms with small cardboard counters bearing the proper numerals and see how easy it is to apply Karp's algorithm to achieve the minimum. Of course, you can make up endless similar tasks, altering the variables k, c and b as you please and shuffling the people any way you like in the building.

If one or more floors have $u_k = 0$ (that is, no one on that floor or below wants to go above that floor), yet some higher floor has u_k greater than zero, the building becomes divided into disconnected regions. The minimum is achieved by handling each region separately according to Karp's algorithm, then piecing together the individual schedules. This procedure increases the number of unit trips by twice the number of floors that must be passed even though they have $u_k = 0$. A little experimenting on buildings with one or more $u_k = 0$ floors below the top one will make clear why that is so. It amounts to the fact that the elevator has to make special trips upward to take care of all higher disconnected regions and then return to the bottom.

Our third elevator problem, discussed on pages 374 through 376 in Knuth's third volume, is based on results obtained by Robert W. Floyd while he was working on efficient ways to rearrange



Robert W. Floyd's elevator problem

records in a magnetic-disk file. This time instead of minimizing distance we want to minimize the number of stops required by the elevator to complete the sorting. Floyd was able to establish a nontrivial lower bound, but no general algorithm is known that achieves the best possible results, except of course a brute-force trial of all possible elevator schedules.

Consider a building where the number of floors, the number of people to a floor and the elevator capacity are each six [*see bottom illustration on opposite page*]. One of Knuth's exercises is to sort the 36 people correctly by starting and ending the elevator on the first floor and to do it in no more than 12 stops. I shall give Floyd's solution next month.

Floyd's method of computing the lower bound is too complicated to explain here, but for this problem it gives nine stops. Even in this simple case it is not known whether there is a solution in nine, 10 or 11 stops. If any reader sends in a solution that is better than 12 I shall report it in a later column.

Our final problem is from Kobon Fujimura's latest Japanese puzzle book, *Dialogue about Puzzles* (Tokyo, 1971; there is no English translation), which he coauthored with Michio Matsuda. Chapter 3 is devoted to an elevator problem that is a cleverly disguised form of a well-known problem in coding theory. In a building of k floors there are n elevators. Each stops on the top and bottom floors and on exactly m floors in between (always stopping on the same m floors). We wish to determine the minimum number of elevators that will enable a person to go from any floor to any other without changing elevators. For example, suppose a building has eight floors and each elevator stops on top and bottom floors and three floors in between. One schedule for a minimum of six elevators that makes it possible for a person to go directly from any floor to any other floor is shown in the top illustration on this page.

As an introduction to this class of problems, readers are asked to answer the following question. Each elevator in a 10-floor building stops on top and bottom floors and four floors in between. What is the minimum number of elevators that will enable a person to ride from any floor to any other without changing elevators? Fujimura's solution will be given next month.

It is impossible to report on all the mechanical puzzles that are appearing on the market, but a new set of five "Piet Hein Puzzlers" calls for comment, not only because readers of this department are familiar with Piet Hein's poems and inventions but also because the new puzzles are of unusual mathematical interest. They are handsomely designed by Hubley Toys, a division of Gabriel Industries in Lancaster, Pa. I shall describe each briefly:

1. Nimbi. This is a 12-counter version of Piet Hein's nim-type game (introduced here in February, 1958) formerly called Tac Tix. The counters are locked-in, sliding pegs on a reversible circular board so that after a game is played by pushing the pegs down and turning the board over it is set for another game. Little is known so far about the best strategy for the polygonal playing field.

2. Anagog. Here we have a spherical cousin of Piet Hein's Soma cube. Six pieces of joined unit spheres are to be formed into a 20-sphere tetrahedron or two 10-sphere tetrahedrons or other solid and flat figures.

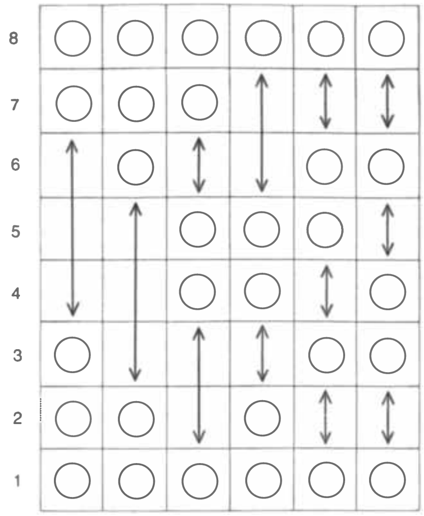
3. Crux. A solid cross of six projecting arms is so designed that each arm rotates separately. One of several problems is to bring three spots of different colors together at each intersection.

4. Twitchit. A dodecahedron has rotating faces and the problem is to turn them until three different symbols are together at each corner. The solution is unique.

5. Bloxbox. W. W. Rouse Ball, discussing the standard 14–15 sliding-block puzzle in his *Mathematical Recreations and Essays*, wrote in 1892: "We can conceive also of a similar cubical puzzle, but we could not work it practically except by sections." Now, 81 years later, Piet Hein has found an ingenious practical solution. Seven identical unit cubes are inside a transparent plastic order-2 cube. When the cube is tilted properly, gravity slides a cube (with a pleasant click) into the hole. Each cube has three black and three white sides. Problems include forming an order-2 cube (minus one corner) with all sides one color, or all sides checkered, or all striped, and so on.

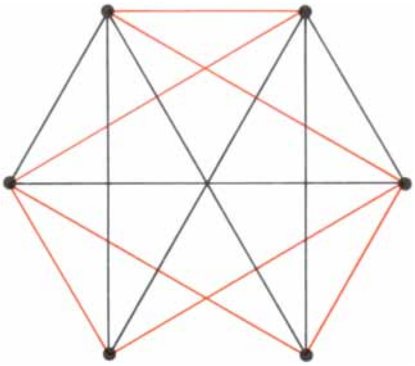
Does the parity principle involved in flat versions apply to the three-dimensional version? And what are the minimum required moves to get from one pattern to another? Bloxbox opens a Pandora's box of questions.

The answer to last month's Sim problem has only one basic position (variants are topologically identical) that allows the game to go 14 moves without a monochrome triangle [*see upper illustration at right*]. The 4-by-5 and 4-by-6 fields for Chomp are won by the unique first moves [*see lower illustration below*].

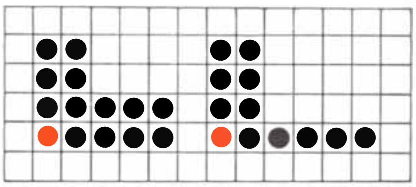


A Japanese elevator problem

David Gale, who invented Chomp, has considered the game on infinite rectangular arrays. Readers may enjoy proving (on the basis of last month's theorems) that the first player wins on n-by-infinity fields (provided that n is not 2) and on infinity-by-infinity squares but loses on 2-by-infinity arrays. A new 500-page edition of *The Guide to Simulations/Games for Education and Training* mentioned last month has been published and is available from Information Resources, Inc., P.O. Box 417, Lexington, Mass. 02173.



Sim game that ends on move 15



Winning chomps

THE AMATEUR SCIENTIST

Molecular models and an interferometer that can be constructed at modest cost

Conducted by C. L. Stong

Certain aspects of nature that lie beyond the direct grasp of the human senses can be represented usefully by physical models, such as the constructions made with balls and sticks to depict the architecture of crystals and molecules. The information conveyed by a model of this kind depends on the meaning that is assigned to its parts. A ball can represent the position of an atomic nucleus or its surface can define the boundary of an atom. The sticks can symbolize the forces that bind atoms together or the direction in which the forces act.

Models of simple molecules can be made inexpensively from kits. The more interesting structures, however, can include hundreds of atoms and can be quite expensive if they are made with commercial parts. A construction technique that enables amateurs to assemble models of giant molecules with inexpensive materials has been developed by Richard E. Goodman, who teaches biology at California State College in San Bernardino, Calif. He has described the technique formally in *The American Biology Teacher* (Volume 32, Number 1, January, 1970). His following discussion is partly based on that article.

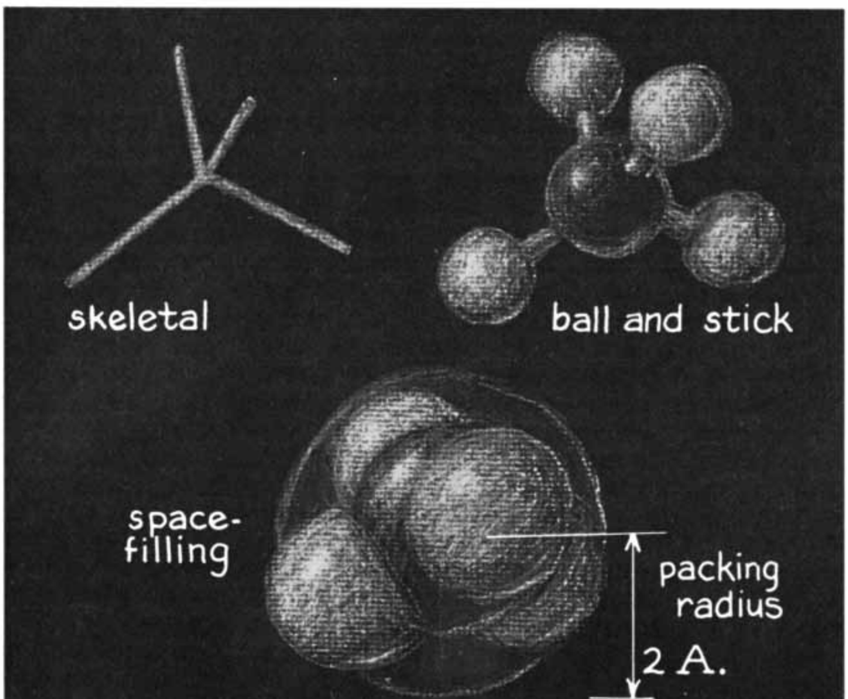
"Models of three basic kinds have been devised to illustrate molecular structures [*see illustration below*]. One employs both balls and sticks. Another consists of sticks without the balls. The third model has balls without sticks. Each type is uniquely capable of illustrating a selected aspect of molecular structures.

"The ball-and-stick model is conventionally used to represent the structure of crystals. The center of each ball symbolizes the location in space of the atomic nucleus. The balls are supported by slender rods in the form of a lattice that matches the configuration of the crystal. The internal structure of the model, including the planes in which the atoms lie, is clearly visible and can be examined in detail. Ball-and-stick models accurately represent the angles at which atoms are bonded to one another. On the other hand, such models do not show the relative sizes of the atoms or the volume of space they occupy.

"Constructions made exclusively of sticks are known as skeletal models. They are valuable for depicting the dimensions of molecules, particularly those that include atoms of carbon. A straight stick symbolizes a chemical bond between two atoms; its free ends mark the locations of the nuclei. The length of the stick is proportional to the internuclear distance. Molecules are represented by two or more sticks joined at points representing the nuclei of the atoms. The angles at which the sticks are joined conform to the directions through which the atom's bonding forces act.

"The skeletal model of a methane molecule, for example, consists of four sticks that meet to form tetrahedral angles of about 110 degrees. The length of the sticks is proportional to the internuclear distance between a hydrogen atom and an atom of carbon. The size of a stick model scaled in this way is proportional to the size of the molecule the model represents. These models also represent accurately the possible spatial configurations of a molecule and therefore can be used in research. For example, skeletal models can be assembled to predict the probable size and spatial configuration of a proposed molecule to be synthesized in the laboratory.

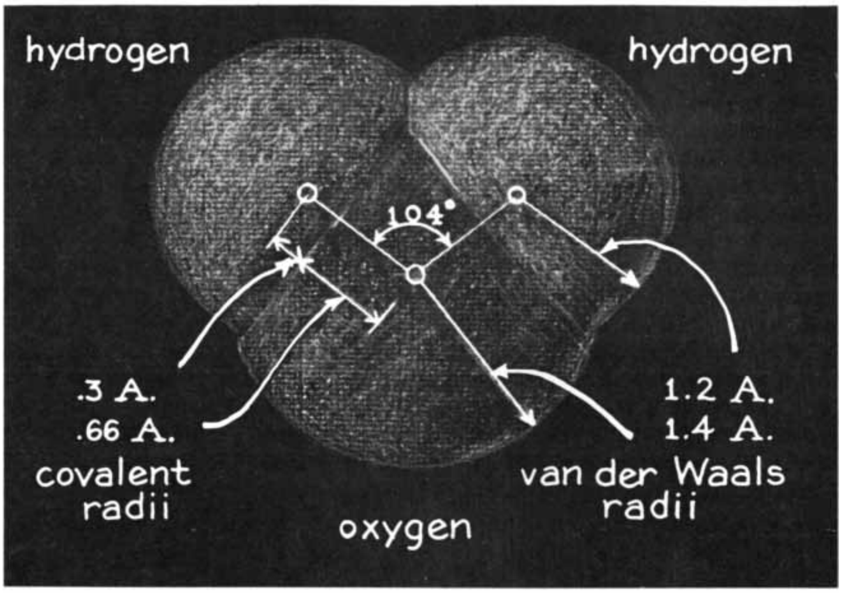
"Constructions made exclusively of balls are called space-filling models. The surface of a ball represents the boundary of a free atom in space. (Such an atom is not attached to neighboring atoms by bonding forces.) The radius of the unbonded ball is a measure of the



A molecule of methane represented by three types of model

effective size of the atom and is known as the van der Waals radius.

"In models of this kind the bonding forces and the angles through which they act are symbolized by cutting away part of the spherical surface to create a facet, or flat surface. For example, a sphere that represents an atom of bonded hydrogen has one facet, which symbolizes the single bond that hydrogen can make with a neighboring atom. A sphere representing an atom of carbon has four facets. Straight lines projected at right angles through the centers of the facets would meet in the center of the ball to form tetrahedral angles of 110 degrees. The depth of the facets varies inversely with the covalent radius of the atom: the distance between the nucleus and the boundary of a bonded atom as measured in the direction of the bonding force. The covalent radius is always shorter than the van der Waals radius.



Richard E. Goodman's space-filling model of the water molecule

"Space-filling models are customarily used to illustrate the structure of biological compounds because they present the viewer with a clear picture of the shape and volume of the molecule: its space-filling property. Moreover, they display precise bond angles. Their dimensions can be conveniently scaled to conform to both the interatomic distances and the van der Waals radii.

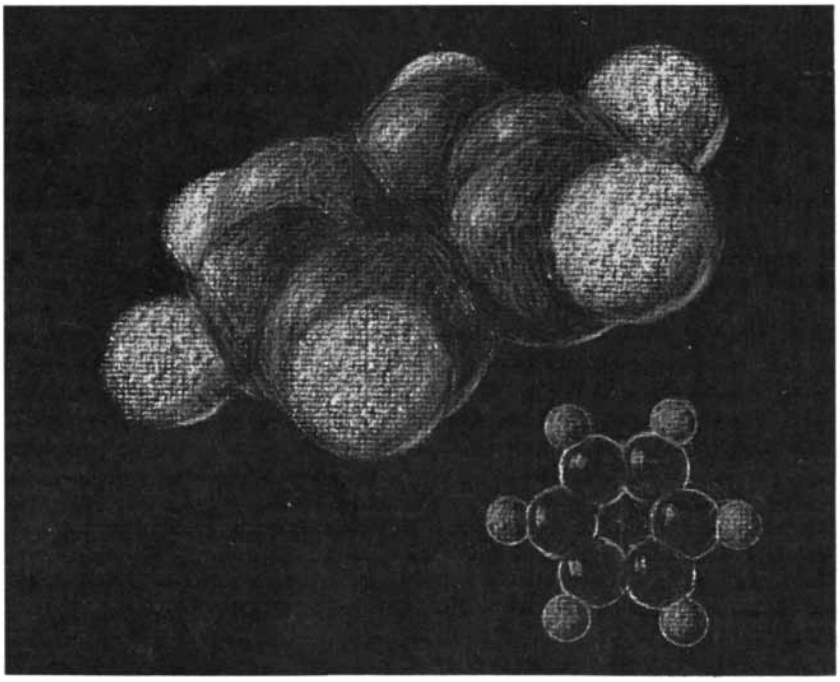
"Various materials are available from which models can be made inexpensively. The plastics industry manufactures spheres in sizes that range from a fraction of an inch to several inches. Some workers prefer to mold their own spheres from epoxy and polyester resins or plaster. Others use methacrylate spheres, which are also available commercially.

"Material for rods and sticks can range from copper tubing and wooden dowel stock to plastic soda straws and chenille pipe cleaners. It takes little ingenuity to devise jigs for cutting facets on balls at accurate bond angles. Great accuracy is not required, however, in models used for demonstrations or exhibits because small errors in bond angles, covalent radii and van der Waals radii are not serious or even noticeable, particularly in models of giant molecules.

"To make my own space-filling models I use balls of expanded polystyrene and join them with connectors made from chenille. The balls could be cemented together, of course, but this would prevent atomic groups from rotating around their bonds and would preclude the possibility of demonstrating alternate structural configurations with the model. Balls of expanded polystyrene are inexpensive. At the time I bought my supply they were priced at about $11 per 1,000 by the Plasteel Corporation (26970 Princeton, Inkster, Mich. 48141). This firm also markets chenille and a coloring kit at reasonable prices.

"I wanted to make several dozen models, including an assortment of giant molecules such as deoxyribonucleic acid, or DNA. To conserve time and effort I devised a few shortcuts that involved the sacrifice of some accuracy in exchange for the desired economies. The resulting models are not of research quality, but they are more than adequate for making demonstrations.

"I used spheres of 1/2-inch diameter to represent hydrogen atoms and spheres of 3/4-inch diameter to represent all other atoms. In addition to carbon the atoms that occur frequently in biological compounds include hydrogen (van der Waals radius 1.2 angstroms), nitrogen (1.5), oxygen (1.4), phosphorus (1.9) and sulfur (1.85). The ratio of the diameters of the small spheres to the large spheres is .67, whereas the ratio of the van der



Benzene modeled by the space-filled method

Waals radius of hydrogen to that of the other atoms ranges from .63 to .86. The maximum error introduced by using 3/4-inch balls for all atoms except hydrogen is therefore not serious. The scale of approximately .22 inch per angstrom allows the construction of giant-molecule models of convenient size.

"As the first step in building a model I make the required atoms by sandpapering flat surfaces on the styrofoam balls to depths that represent the covalent radii at the correct bond angles. The angle between a pair of facets is supplementary to the desired bond angle. For example, an atom of oxygen unites two atoms of hydrogen at a bond angle of 104 degrees to form a molecule of water [*see top illustration on preceding page*]. The supplementary angle between the facets is equal to 180 degrees minus 104 degrees, or 76 degrees.

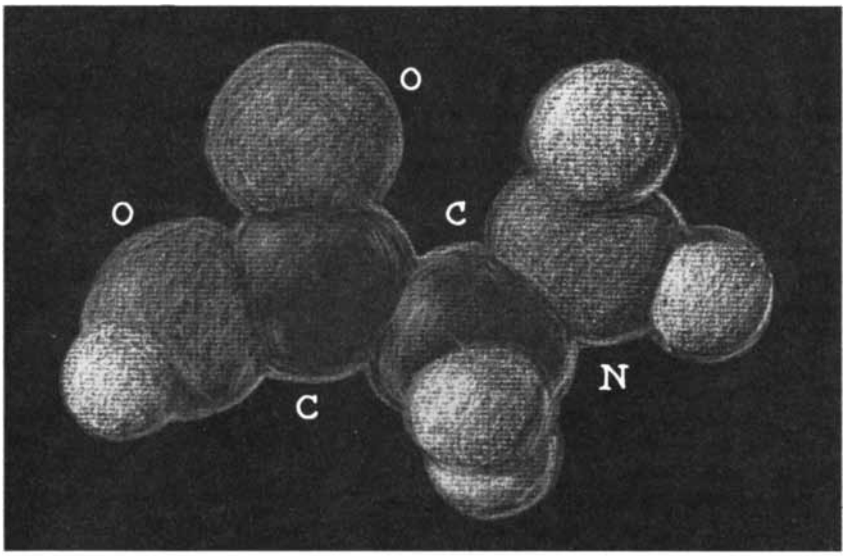
"To make the facets I use a 3/8-inch electric drill fitted with a disk sander. The depth to which the plastic is sanded and the bond angles between facets are judged by eye. Balls that represent hydrogen atoms are held in contact with the sandpaper by needle-nosed pliers. The larger balls are held by hand.

"Atoms frequently used in models of biological compounds are listed in the accompanying table [*page 114*] together with the required number of facets and the covalent radii. Bond angles of the tetrahedral carbon atom should be about 110 degrees, of the dihedral oxygen atom about 104 and of the sulfur atom about 102.

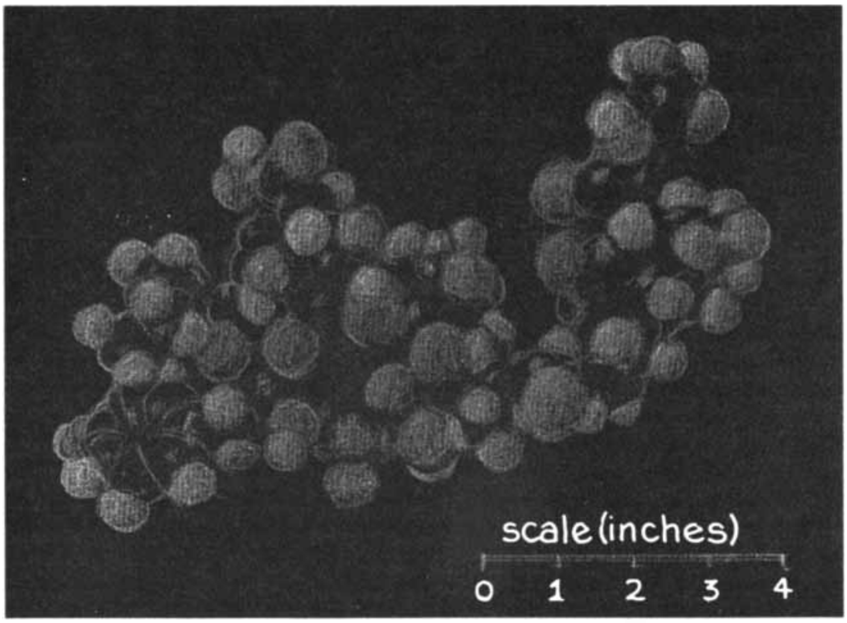
"Atoms other than those of hydrogen must be labeled in some way because they are all made with balls of the same size. I identify the various chemical elements by means of a color code that is in common use: white for hydrogen and phosphorus, black for carbon, red for oxygen, green or blue for nitrogen and yellow for sulfur. To apply the color I push a toothpick into the ball and immerse the plastic in an alcohol-soluble dye. I then invert the toothpick and stick the free end into a slab of styrofoam to support the ball in air until it dries.

"To assemble the model I push a piece of chenille into the center of a facet to a depth of about a quarter of an inch. With wire cutters I snip the chenille at a point about a quarter of an inch from the facet. The mating facet of the ball to be joined is pushed into the projecting chenille. The fibers of the chenille fold smoothly against the wire spine as the connector passes into the plastic but act as barbs when they are pulled in the other direction. The bond is surprisingly strong, but the assembly can be pulled apart without damaging the plastic.

"I have illustrated the structures of many compounds with the models. Sequences of models have also been constructed to depict enzymatic reactions and even entire metabolic pathways. Models of polypeptides, or chains of amino acid units, are easy to construct. The low density of expanded plastic and the small size of the spheres result in models of giant molecules that are easy to handle, store and transport. A model of DNA consisting of 25 base pairs was assembled from its constituent atoms in a single evening. The resulting structure, which shows the celebrated double helix, was supported by scaffolding fabricated with D-Stix, a construction toy."



A model of glycine



A model of ocytocin, a polypeptide hormone

In a few of its several forms the interferometer serves as a convenient instrument for observing and recording photographically local variations in the density of a gas, such as disturbances caused in air by the heat of an open flame or the flow of air around an object in a small wind tunnel. Essentially the instrument splits a source of light into two beams of parallel rays, recombines the beams at a distant point and focuses the converging rays on a screen. The optical parts of the instrument, which consist of mirrors and lenses, can be ad-

THE SEVEN DEADLIEST CRIMES AGAINST YOURSELF!

Are You Guilty of Any of Them?

1 **How many times in the last 24 hours did you risk a heart attack or an ulcer — simply because you didn't know a little fact about proper breathing?**

2 **When was the last time you snapped at an imagined insult — and lost a true friend?**

3 **Are you missing respect and popularity — by a hairsbreadth — because you're overlooking a simple principle of human dynamics?**

4 **Is insomnia ruining your life — because you're omitting a 20-second, nightly routine?**

5 **Are you starving yourself sensually — without even knowing it? Without knowing what to do about it?**

6 **Do you cause secret misery to loved ones — when you're actually trying your hardest to help them?**

7 **Are you leaving yourself open to daily tensions that can snowball into disastrous "blow-ups"?**

LAURA HUXLEY, the authoress, is one of the most remarkable women of our time. Wife of world-famous ALDOUS HUXLEY, she is renowned as a successful therapist.

If you're "guilty" of even one—you can turn the tables of tensions and anxieties with the help of this extraordinary new book

Laura Huxley, the author, is one of the most remarkable women of our time. Writer, musician, wife of world-famous Aldous Huxley, she is renowned as a dramatically successful therapist.

Her life-long crusade to increase human happiness and creativity has taken her to virtually every major civilization and culture of the world.

From a Lifetime of Discovery— Potent Secrets For You

Laura Huxley has mastered the Buddhist secrets of tranquility. She has applied the healing power in the words of the Bible. She has identified and traced the life giving principles that lead from the sages and philosophers of the ages to the psychotherapeutic advances of our own scientific tradition.

Through her extensive research and experimentation she has helped hundreds who have come to her for help and counsel. She has worked with equal enthusiasm and success in the mansions of millionaires and in the wards of hospitals. (One of her most gratifying triumphs occurred at a famous U.S. psychiatric clinic.) But in her own words, "My greatest satisfaction has always come from showing perfectly normal people how to make their lives richer than they ever dreamed possible."

And now in YOU ARE NOT THE TARGET, she reveals to you the essence of the exciting program which has made her such a sought-after counselor.

Why You'll Love Every Minute of Laura Huxley's Revolutionary Program

- **It's completely private.** You can do all of it all by yourself.
- **It's fun** . . . exciting, exotic downright unconventional. (Who ever heard of "smacking a Tether Ball"? But try it and you'll reap a harvest of relaxation that will brighten all your activities.)
- **It costs nothing.**
- **It will strengthen and beautify your body**—effortlessly. (Even as it gives you new happiness and peace of mind.)
- **It begins to work immediately**—the very moment you read Page One.

Here are just a few secrets of successfully living in YOU ARE NOT THE TARGET:

- How to conquer love-starvation.
- Sensual Pleasure. (You may never have really experienced it before.)
- A simple body rhythm—more beneficial than strenuous exercise. (This is the famed isotonic routine endorsed by the U.S. Olympic team.)
- "Bubble Freedom"—a giddy technique to clear your thinking. Ingredients: one pan of water and plenty of fun.
- How to start liking yourself. (You'll be surprised how much more others will like you.)
- The story of a beautiful girl haunted by a ghost—that will help vanquish the ghosts in *your* life.
- How to deal with pain.
- Something you can do anywhere, anytime—to help stop wrinkles.
- The fantastic value of silence.
- Delicious, *natural* substitutes for chemical tranquilizers.
- SPECIAL FOR YOUR CHILD—Some common-sense advice that can do wonders for his self-confidence at school—plus a tip on how to get him to eat carrots—or any other food *he thinks he hates.*
- When a headache strikes—strike back! And throw away habit-forming drugs.
- How to feel even better at the end of each working day than you did at the beginning.
- Three things you must never say to your children—but almost everyone does.
- Miraculous water therapy—for sound sleep. It takes 20 seconds. It doesn't disturb your nightly routine.
- Master rules that cover what to eat and what *not* to eat—to keep healthy.
- How one dynamic executive in an "ulcer job" never loses his peace of mind and never gets heartburn.
- How to breathe and stretch—for magic dividends in health and beauty.
- SPECIAL HELP—FOR THE SERIOUSLY TROUBLED—if you are one of the lucky people who have learned to live successfully in this modern world, why not share this great gift with a troubled friend or loved one? Introduce him or her to the therapies in this book. (You'll probably find that the counsel you've been meaning to share is strikingly close to the principles in YOU ARE NOT THE TARGET.)
- How to combat your "enemies"—in a way much more satisfying than slugging them—or snubbing them.
- The judo of dealing with tension.
- Three minutes that can end auto accidents.
- How to add glamour to all your cooking all the time.
- How to get rid of compulsion, fear, embarrassment and envy.
- How to cut down on smoking—painlessly and naturally.
- A simple, effective device that actually recharges your home with exhilaration. (The ancient Persians and the Moors discovered it—science has verified it.)
- The one nonsense syllable that makes sense. Use it the next time you get into a mental rut.
- Secrets of confidence for daily living—discovered by a leading anesthetist in the operating room.
- How to delve into your sub-conscious—and release untapped resources.
- A mental approach to improve your figure and your posture—without moving a muscle or counting a calorie!
- Techniques for unblocking your creative capacity. (This alone could mean the difference between success and just "getting by" in your career.)
- A reprieve for your heart—if you follow some simple advice.
- What sleeping with a loved one can mean.
- Ageless lessons from a Hindu God—to his beloved.
- To hell with caution! Dedicated to those over 40.
- How to be human—in an inhuman world.
- *and much, much more!*

Can you imagine how rich your life would be if you were freed of your heaviest emotional burdens? You don't have to imagine it. This book will show you how to live it.

JAY NORRIS CORP.

Dept. BGK-4,
25 W. Merrick Rd., Freeport, N.Y. 11520

Serving Satisfied Customers for over 25 Years

30-DAY MONEY-BACK GUARANTEE
MAIL NO RISK COUPON TODAY!

JAY NORRIS CORP., 25 W. Merrick Road
Dept. BGK-4, Freeport, N.Y. 11520

Please rush me_____copy(ies) of "YOU ARE NOT THE TARGET" @ $6.95 complete. We pay postage.

☐ SAVE! ORDER TWO for only $12.95 complete.

Enclosed is ☐ check or ☐ money order for $_____

(N.Y. residents add sales tax.)

PRINT NAME__________
ADDRESS__________
CITY__________
STATE__________ZIP__________

© Jay Norris Corp., 1972

justed so that light waves in one beam of parallel rays fall slightly out of step with those in the other beam. When the beams are recombined and projected on the screen, the crests of the waves will coincide in some areas of the screen to form bright patterns. In other areas the crests of the waves in one beam will coincide with the troughs of the waves in the other beam; the screen appears dark in these areas.

When the instrument is appropriately adjusted, the screen displays a gridlike pattern, being crossed by fringes, or alternately dark and light bands, that are uniformly spaced and parallel. If either of the split beams traverses a gas of nonuniform density in an instrument so adjusted, the form of the grid is altered to a pattern that corresponds to the density variations. Interferometers of conventional design must be made with parts of good optical quality because the projected image displays imperfections in the lenses and mirrors as well as nonuniformities in the density of a gas. Instruments of this quality are priced beyond the reach of most casual experimenters. Chris F. Bathurst of Invercargill in New Zealand has devised a scheme for making a useful interferometer from scraps of ordinary plate glass and with lenses of the kind used in inexpensive magnifying glasses. He discusses the construction as follows:

"My instrument is a variation of the cyclic interferometer, so called because the test and reference beams of light traverse a triangular path in opposite directions. The unusual feature of the instrument is a lens that is inserted in the base of the triangular path. The lens can be adjusted to minimize the effect of imperfections in the optical parts.

"Rays of light from a conventional source are focused into a converging cone that falls on a beam splitter in the form of a partially silvered mirror [*see illustration on opposite page*]. The apex of the transmitted cone passes through the test section of the instrument—the vessel that contains the gas specimen. The cone is deliberately located in a region of the vessel that is known to contain stable gas.

"The rays proceed through the test section as a diverging cone and are reflected by a mirror to the lens in the base of the triangular path, as shown by the colored rays in the illustration. The lens bends the rays into a parallel bundle that proceeds to the second mirror, which reflects the bundle to the beam splitter, thus closing the triangular path. Half of the light is reflected to the source by the beam splitter and is lost. The beam splitter transmits the remaining half to a field lens that focuses the rays at the viewing position. Essentially these rays function as a reference beam with which rays that are disturbed in the test chamber are compared.

"Converging rays from the source that are reflected by the beam splitter traverse the triangular path in the opposite direction, as shown by the black rays in the illustration. Following this reflection they impinge on the second mirror and are reflected to the lens in the base of the triangular path. The lens transforms the diverging cone into a beam of parallel rays. After subsequent reflection by the first mirror the beam floods the full area of the test section with parallel rays. The direction and the velocity of the rays will be altered more or less by local variations of density in the gas.

"Rays that emerge from the test section return to the beam splitter. Half of the light proceeds through the beam splitter to the source and is lost. The remaining rays are reflected by the beam splitter, interfere with rays of the reference beam and proceed through the third lens to the focal plane at the viewing position. If the gas in the test section is of uniform density, the instrument can be adjusted to project a pattern of fringes that are reasonably parallel and straight even though the surfaces of the optical elements are imperfect. The effects of optical imperfections can be reduced about 80 percent. The residual imperfections in the fringe pattern of the instrument described here are about seven fringes in a field of view three inches square.

"None of the lenses I used is of high quality. All are symmetrical. Lens 1 (focal length 10 inches, diameter four inches) is from a magnifying glass. Lens 2 (focal length 42 inches, diameter four inches) and lens 3 (focal length 25 inches, diameter four inches) were made by a local craftsman with a machine that is normally used for grinding and polishing spectacle lenses. Lens 4 (focal length 5½ inches, diameter 1¼ inches) is of relatively small diameter and reasonably good optical quality. Most of the light passes through it close to its axis. Hence the lens introduces little distortion.

"The green filter (Wratten No. 64) restricts the light in the image to monochrome. I use it for isolating the 5,461-angstrom line emitted by a mercury lamp. The filter can be omitted if a laser is used as the light source.

Atom	Bond configuration	Number of facets	Approximate covalent radius (angstroms)
Hydrogen	H–	1	.3
Carbon (tetrahedral)	–C– (with single bonds above and below)	4	.8
Carbon (aromatic)	C (double bond above, single bonds to either side below)	3	one at .8, two at .7
Oxygen (dihedral)	⁄O⟍	2	.7
Oxygen (double-bonded)	O=	1	.6
Nitrogen (tetrahedral)	–N (with single bonds above and below)	3	.7
Nitrogen (aromatic, dihedral)	N (double bond at upper right, single bond at lower right)	3	.6
Nitrogen (aromatic, trihedral)	–N (single bonds at upper and lower right)	3	.7
Phosphorus	–P= (with single bonds above and below)	4	three at 1.1, one at 1
Sulfur	S (single bonds at upper and lower right)	2	1

Data on atoms common in biological molecules

"All mirrors and the windows of the test section were made of ordinary plate glass 3/8 inch thick. The reflecting surfaces are films of aluminum deposited in a vacuum. A series of partially coated beam splitters was made to determine the optimum thickness of the aluminum in terms of reflection. I found that the thickness of the film is not crucial. The metallized coating must be on the surface of the beam splitter that is nearest the test section.

"The mirrors are supported at three points by mechanical mounts that were improvised from available materials. Small pieces of cardboard separate the metal clamps from the glass. The windows of the test chamber are sealed with gaskets of cork. Two micrometer screws, which function as adjustments, rotate the mirror mounts about the horizontal and vertical axes in the plane of the mirror.

"In addition to these adjustments the mounting of lens 2 must be adjustable in both directions parallel to the optical axis and at a right angle to the axis. I did not fit this mounting with screws for making translational adjustments, although they would be convenient. The base of the instrument is made of 1/4-inch steel channel sections and measures three inches in width by 1½ inches in height. Photographs of fringe patterns were made with a 35-millimeter camera.

"The adjustment procedure is somewhat tedious, as it is with most interferometers. My instrument can be aligned and adjusted in from 10 to 30 minutes, depending on the experience of the operator. After the mountings of the optical elements have been assembled to the base remove the beam splitter and lenses 2 and 3. Put a rectangle of white cardboard at the position normally occupied by lens 3. Move lens 1 toward the light source to the point at which the lens projects a light beam of substantially constant width. Align the mirrors to the position at which the beam illuminates the screen.

"Adjust lens 1 and mirrors 1 and 2 to focus a sharp image of the light source on the center of the screen. Replace the beam splitter. Two images of the source should appear on the screen. Adjust the beam splitter to the position at which the two images coincide exactly.

"Move lens 1 toward the light source. A pattern of light and dark fringes should appear on the screen. If fringes do not appear, return to the previous step and again attempt to adjust the beam splitter to the exact point at which the images of the source coincide.

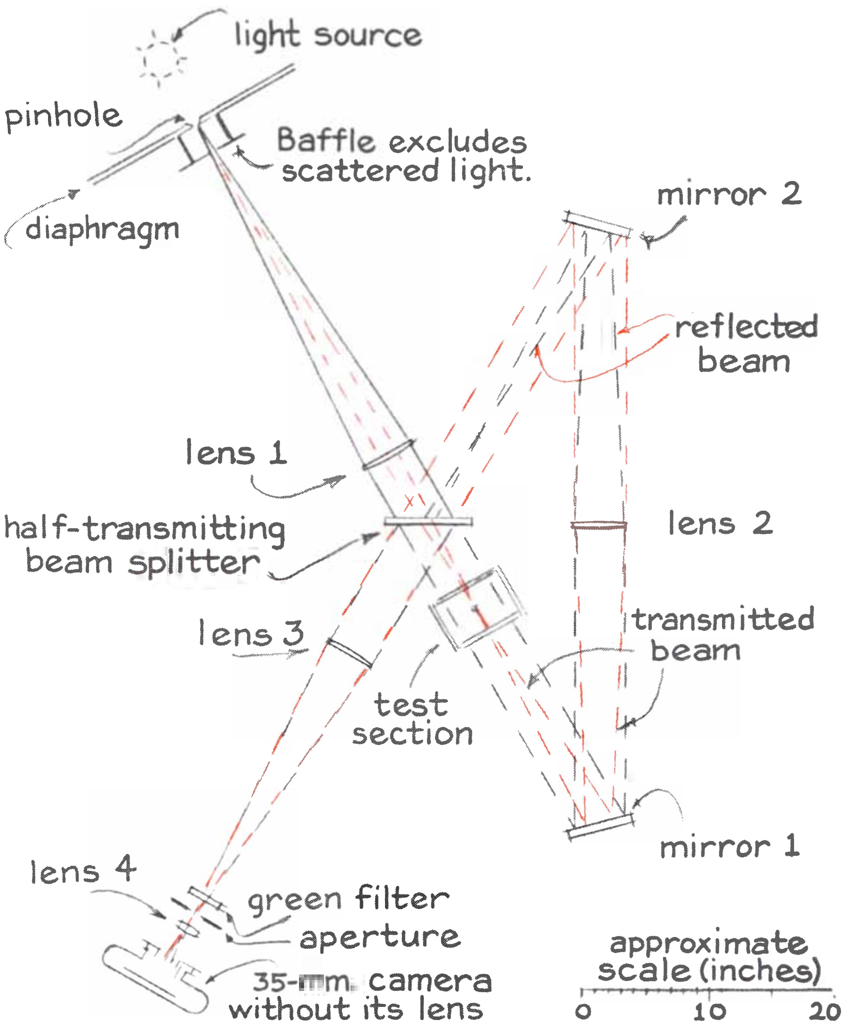
"Look for the fringes. Repeat the procedure until they appear. The operation calls for patience until you acquire experience. After the fringes appear adjust the beam splitter to the position at which you observe the minimum number of fringes.

"Replace lens 2 and move lens 1 toward the beam splitter to the point at which the apex of the cone of converging light rays lies inside the test section. The location of the apex within the test section can be altered by shifting lens 1 up and down or sideways. Avoid regions where the apex would interfere with the phenomenon under study.

"Next, move lens 2 along its axis to the position at which fringes fill the field of view. If the fringes do not appear, return to the critical step in which the images of the light source were made to coincide by manipulating the beam splitter. When the field of view eventually fills with fringes, manipulate lens 2 and the mirror mounts to remove as many fringes as possible. Adjust either of the mirrors to the position at which a desired pattern of fringes, such as a grid, appears in the field of view.

"Finally, transfer the image to the focal plane of the instrument by replacing lens 3, the field lens. Adjust the position of aperture near the camera to intercept the halo of stray light that surrounds the converging rays. The position of the focal plane of lens 4 can be located with a screen of white cardboard.

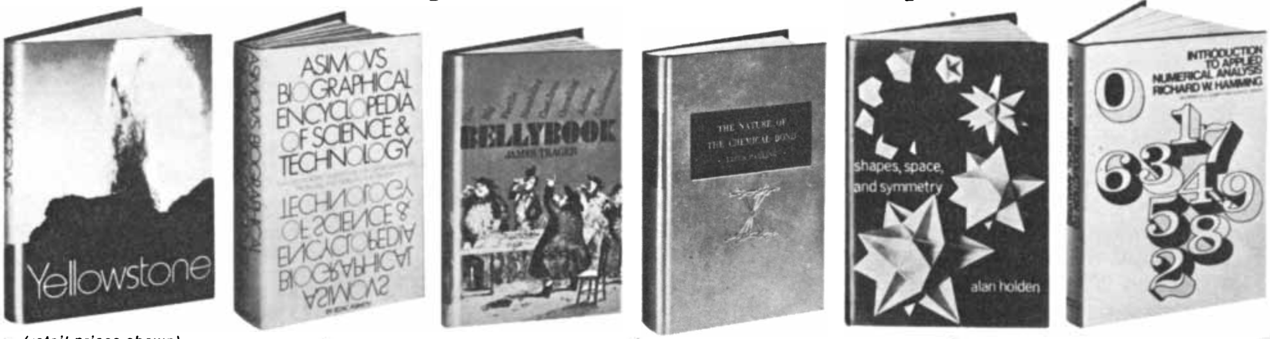
"To make sharp photographs of fringes that arise from phenomena in the test section place the film of the camera in the focal plane of the instrument. I have used light sources of three kinds. A frosted incandescent lamp can be located behind lens 1 for adjusting the camera to the position of sharpest focus with respect to phenomena in the test section. All other adjustments have been made with a 250-watt, high-pressure mercury arc lamp. Photographic exposures are made by replacing the mercury lamp with a xenon flash lamp of the kind available from dealers in photographic supplies."



Optical path of Chris F. Bathurst's cyclic interferometer

How many of these Best Sellers belong in your scientific library?

Get 3 that you don't have—for only $3.95.



(retail prices shown)

74010. THE RESTLESS EARTH. *Nigel Calder.* Experts at the very frontiers of geology report on the exciting findings that are revolutionizing man's thinking about the earth. Profusely illustrated. **$10.00**

61340. MATHEMATICAL STATISTICAL MECHANICS. *Colin J. Thompson.* A mathematical approach to classical statistical mechanics with important applications to biology. **$12.75**

45490. ENVIRONMENTAL GEOLOGY. *Peter T. Flawn.* The relationship between man and natural earth processes is keyed to the improved use and control of the environment. **$14.50**

62460. THE METAPHORICAL BRAIN: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory. *Michael A. Arbib.* Introduces a new concept of the brain as an action-oriented, distributed computer, based on the latest experimental research and from recent advances in the design of robots. **$14.95**

63910. THE NATURE AND FUNCTION OF SCIENTIFIC THEORIES. *Edited by Robert G. Colodny.* Six scholars analyze the contemporary status of scientific theories. A striking and original contribution to the philosophy of science. **$12.95**

37530. CHEMICAL EVOLUTION. *Melvin Calvin.* Nobel laureate Melvin Calvin's major work on the biochemical evolution of life. Emphasizes the transition period between the appearance of nonliving molecules and those which must be called "living." **$9.00**

82640. TECHNOLOGY IN THE ANCIENT WORLD. *Henry Hodges.* Technical survey of the tools and processes of early civilizations. A delightful and useful volume for the professional and nonprofessional. **$10.00**

63870. NATURE IN MINIATURE. *Richard Headstrom.* A fascinating, month-by-month look at nature's most subtle creations — with 561 pen and ink drawings. "A beautiful volume."—*Choice* **$10.00**

39891. THE COMPUTER FROM PASCAL TO VON NEUMANN. *Herman H. Goldstine.* One of the world's leading mathematicians describes graphically the development of the computer. A highly personal account by one of the "fathers" of this twentieth-century miracle. **$12.50**

34610. ASIMOV'S BIOGRAPHICAL ENCYCLOPEDIA OF SCIENCE AND TECHNOLOGY. New Revised Edition. *Isaac Asimov.* The lives of 1,195 of the world's great savants from ancient Greeks to molecular biologists. **$12.95**

78340. SHAPES, SPACE, AND SYMMETRY. *Alan Holden.* By word and picture, explains, analyzes, and provides amazing insights into hundreds of mathematical solids ranging from the tetrahedron to the trapezoidal hexacontahedron. **$11.00**

36010. THE BIG, FERTILE, RUMBLING, CAST-IRON, GROWLING, ACHING, UNBUTTONED BELLYBOOK. *James Trager.* Startling insights into the machine that is the human body; an enormous compendium covering diet, digestion, metabolism, nutrition, vitamins, food safety, more. **$15.00**

39960. COMPUTER HARDWARE THEORY. *W. J. Poppelbaum.* Covers every aspect of the theory used in designing computers and their input/output equipment. A valuable and complete guidebook. 730 pages. **$16.95**

40040. THE CONCEPTUAL DEVELOPMENT OF QUANTUM MECHANICS. *Max Jammer.* "A book to read and savor, not once, but several times." —*The Journal of Chemical Education.* **$12.50**

49350. FUNDAMENTALS OF PHYSICS. *David Halliday and Robert Resnick.* A complete updating and revision of the modern classic that has sold more than one million copies. Stresses theory rather than technology and is selective rather than encyclopedic. **$13.95**

34620. ASTRONOMY: Fundamentals and Frontiers. *Robert Jastrow and Malcolm Thompson.* Stimulating new book that breaks with tradition. By the author of the best-selling *Red Giants and White Dwarfs* and Director of the Goddard Institute for Space Studies. **$12.50**

50380. GLACIAL AND QUATERNARY GEOLOGY. *Richard Foster Flint.* A just-published masterwork on climatic change and glaciation over the last ten million years. Required reading for anyone interested in geology, geophysics, meteorology, botany, and conservation. *Counts as 2 of your 3 books.* **$24.95**

40110. THE CONDENSED CHEMICAL DICTIONARY. Eighth Edition. *Edited by Gessner G. Hawley.* Revised, expanded, updated and full of new features. Contains 18,000 entries, more than 1,000 pages. *Counts as 2 of your 3 books.* **$27.50**

56030. INTRODUCTION TO APPLIED NUMERICAL ANALYSIS. *Richard W. Hamming.* The basics of computational mathematics, with special emphasis on techniques useful with digital computers, by the Head of Bell Labs' Computer Science Research Department. **$14.95**

46590. EVOLUTIONARY OPERATION. *George E. P. Box and Norman R. Draper.* Evolutionary Operation, an exciting process improvement technique pioneered by the authors utilizes the concepts of natural selection and the tools of mathematical statistics to maximize industrial efficiency. **$11.00**

44350. ELEMENTS OF CARTOGRAPHY. Third Edition. *Arthur H. Robinson and Randall D. Sale.* Just published, this completely revised edition includes a wealth of new material, especially on techniques of air-mapping, and the most recent experiments in design and reproduction. **$11.95**

67661. THE PENGUIN DICTIONARY OF SCIENCE. Fourth Edition. *E. B. Uvarov and D. R. Chapman. Revised and Edited by Alan Isaacs.* Complete reference tool covering the biological and physical sciences in one comprehensive volume. **$11.95**

48531. FINITE MATHEMATICS WITH APPLICATIONS. *Abshalom Mizrahi and Michael Sullivan.* Brilliant work that puts such tools as probability, linear programming, matrices, directed graphs, Markov chains, game theory and statistics within the reach of the non-specialist. **$10.95**

78190. SEXUAL SELECTION AND THE DESCENT OF MAN. *Edited by Bernard Campbell.* Newly published and remarkable evaluation of the scientific impact of Darwin's ideas, by ten outstanding authorities in different disciplines. **$14.75**

63950. THE NATURE OF THE CHEMICAL BOND. *Linus Pauling.* Third edition of this science classic. All chapters expanded to include changes in such matters as bond-energies and bond-lengths. **$17.50**

88120. YELLOWSTONE. *Ann and Myron Sutton.* Traces the history and explores the future of this National Park now celebrating its hundredth anniversary. Beautifully illustrated. *Counts as 2 of your 3 books.* **$25.00**

Take any 3 books (values to $49.45) for only $3.95

with a short trial membership in

The Library of Science

if you will join now and agree to take only 3 more books during the next twelve months

MEMBERSHIP APPLICATION 2-36Y

The Library of Science, Riverside, N.J. 08075

Upon acceptance of this order, please enroll me as a member and send the 3 books I have indicated. Bill me only $3.95, plus postage and handling. If not delighted, I will return all books within ten days and this membership will be cancelled.

As a member, I need accept only three more selections during t e next 12 months at reduced member prices, plus postage and handling. Savings range up to 30% and occasionally even more. I understand that I will receive free advance Reviews which fully describe each month's Main Selection and Alternates. If I wish the Main Selection, I need do nothing and it will be sent automatically. If I prefer an Alternate—or no book at all—I need only return the convenient reply card you send me by the date specified. Send no money. Members are billed when books arrive.

3 books for $3.95. (write in numbers)

[ ] [ ] [ ]

(Some expensive books count as 2 choices.)

Name__________

Address__________

City________ State________ Zip________

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.)

BOOKS

A classic reexamined: Robert Merton's study of science in the 17th century

by I. Bernard Cohen

SCIENCE, TECHNOLOGY AND SOCIETY IN SEVENTEENTH-CENTURY ENGLAND, by Robert K. Merton. Harper & Row ($2.75).

For more than 30 years Robert K. Merton's great inquiry into science, technology and society in 17th-century England has been more honored than read. The reasons are both technical and stylistic. On the technical side this book-length study was originally published in a journal (*Osiris*) of limited circulation and so was available to nonsubscribers only in major libraries. On the stylistic side Merton's work carried a substantial armament of graphs and tables. in a day when few historical analyses (even of science) were given quantitative expression. Moreover, the presentation had a somewhat ambiguous character: was it a historically based sociological analysis of science or a sociologically oriented view of an episode in history?

If Merton's book was so hard to find (one could hardly buy a copy for one's own library) and so difficult to read in a day when quantitative history was unusual, how did it come to be revered as a classic? Chiefly, I suspect, through some general articles in which Merton expounded certain of his principal findings. Those articles have been reprinted in a number of widely read anthologies and in a collection of Merton's papers that has gone into several editions.

This form of dissemination has concentrated attention on one of Merton's conclusions at the expense of others. It is a conclusion now known as the "Merton thesis." It is that the rise of modern science can be closely linked with the "Protestant ethic" in general and in England with the ascetic part of Protestantism associated with Calvin and the Puritans. This aspect of Merton's work has tended to overshadow another important contribution he brings to the elucidation of science as an institution: his cogent analysis of science and technology, notably military technology.

Today's reader is not put off by numerical or quantitative considerations. Moreover, many sociological studies of science now draw on historical information, and there has grown up a generation of historians of science and historically minded scientists who have become increasingly concerned with science in its relation to the social environment. In the 1970's there are enough prospective readers to justify a reprinting of Merton's book in paperback form.

At the outset Merton addresses himself to general questions concerning the sociology of knowledge. Why are there shifts in the primary foci of intellectual attention from one age to another? At different times, he observes, the focus has been on philosophy, art, religion, literature and ethics. In the past three centuries "the center of interest seems to have shifted to science and technology." Aware that internal developments and pressures within "fields of culture" are partly responsible for the attention given to a field, he is nonetheless attracted by the possibility that "other social and cultural conditions have also played their part."

This hypothesis gives rise to certain fundamental questions: Which social processes are involved in shifts of interest from one division of human activity to another? What, indeed, is the nature of the sociological conditions that are associated with pronounced activity in any of these domains? The questions are far too general for fruitful inquiry, and so Merton decided to illuminate the general problem by the specific case, namely "the sociological factors involved in the rise of modern science and technology."

As a sociologist-historian Merton sets out to analyze the scientific revolution of the 17th century, particularly as it was manifested in England. It would be difficult to find a period and place richer than 17th-century England in material relating to "shifts and foci of interest in science and technology." The century began in 1600 with the publication of William Gilbert's fundamental monograph *De Magnete*, which was perhaps the founding work of modern experimental science. It was a time of great mathematicians—Napier, Briggs, Wallis, Wren, James Gregory, Harriot, Oughtred and above all Newton. It was the age of the great experimentalists, Boyle and Hooke, of the founder of physiology, Harvey, of Ray and Willis in the life sciences and of Flamsteed and Halley in astronomy. The century saw the founding of the Royal Society of London and of its *Philosophical Transactions*, which is perhaps the oldest regularly appearing scientific periodical. The crowning achievement of the century was Newton's *Principia*, setting forth the foundations of dynamics and celestial mechanics.

Dismissing the possibility that the scientific revolution arose from a mere "simultaneous appearance of geniuses," Merton finds a "more plausible explanation" in the "combination of sociological circumstances, of moral, religious, aesthetic, economic and political conditions, which tended to focus the attention of the geniuses of the age upon specific spheres of endeavor." In order to document the shifts in vocational interest in the 17th century Merton made a card for every person appearing in the British *Dictionary of National Biography* who by certain rules could be assigned to the century. Then he sought for information concerning the approximate time when each such person "first evidenced interest in his chosen field or fields." He grouped the determinations into five-year periods and then drew up tables showing the types of interest according to the categories of military service, painting and sculpture, music, drama, poetry, prose, education, historiography, medicine, religion, science, scholarship, law and politics. The results are also given in a series of graphs.

A peak occurs in the category of military service 1641–1645, one causative factor being the establishment of a standing army in England; another peak in the latter 1680's coincides with the campaign of William of Orange. The graph for medicine and surgery shows a rise to a high level at mid-century and a fairly constant high level thereafter. The graph for religion shows "an almost continuous and unbroken decline of interest in the ministry as an occupation," even though "religion remained one of the dominant social forces throughout this period." The curve for science displays "a consistently increasing interest in science in the first half of the century, reaching its peak in the quinquennium 1646–50." It is difficult to disagree with Merton's conclusion that the rise of "science as a social value" in that time and the decline of other vocations may have been closely related to considerations of utility.

Merton next proposes as a hypothesis that "we may expect a lag of ten or twenty years between the time of initial interest and significant scientific productivity." Thus "one would look for the peak of scientific discoveries and inventions in the decade 1661–70." It is significant that this is the decade of the establishment of the Royal Society and the beginning of its *Philosophical Transactions*. Within the sciences the data show a greater vocational concern for the more formal and physical sciences (mathematics, astronomy, physics, chemistry) until about 1685, when there is a decline, but "the biological sciences—natural history, botany and zoology—maintain an approximately constant share of scientific attention save for a small increase in 1691–93."

The facts thus established, Merton proceeds to three major problems of historical sociology or sociohistory: (1) Puritanism and the new science, (2) science and technology and (3) "extrinsic influences on scientific research" and "cultural factors in scientific advance." As a sociologist Merton was heir to Max Weber's doctrine about a special aspect of Calvinist Protestantism, which Weber called "the Protestant ethic" and saw as contributing a rationally based attitude of approval for things of "this world." Merton extends Weber's analysis to science and in particular shows by numerical data that the "Puritans" of 17th-century England made a far greater contribution to the new science than one would expect from their number. He has summarized his findings in the statement that "the combination of rationalism and empiricism which is so pronounced in the Puritan ethic forms the essence of the spirit of modern science."

Puritanism suggested that God was revealed in his handiwork, so that nature could, through science, be a means for appreciating "the wisdom of God in creation" (the title of a book by John Ray). Even more important, says Merton, was a "dominant tenet in the Puritan ethos" that "designated social welfare, the good of the many, as a goal ever to be held in mind." For the Puritan, Merton thus concludes, science "was to be fostered and nurtured as leading to the improvement of man's lot on earth by facilitating technologic invention."

The thrust of Merton's argument is that Protestant England in the 17th century provided a particularly suitable environment for the development of modern science, and Merton seeks to make precise which social and intellectual aspects of that time and place may have been the factors most responsible. He is fully aware, however, that science had also developed in Catholic Italy and France. Obviously Galileo, Descartes and Pascal were not Protestants, although their contemporaries Tycho and Kepler were; Steno (a founder of crystallography) started out as a Protestant in Denmark and ended up as a cardinal in Italy.

To my mind Merton's two most original contributions are his pioneering work in establishing a methodology for the quantitative study of history and his analysis of the relations of science and technology, leading to his explorations of economic, social and cultural interactions with the advance of science. Four of the 11 chapters of the book are devoted to "economic and military influences on the spectrum of scientific investigation." Merton demonstrates convincingly that "the relation between science and economic needs is seen to be twofold: direct, in the sense that certain scientific research is advisedly and deliberately pursued for utilitarian ends, and indirect, in so far as certain subjects, because of their technologic importance, are sufficiently emphasized to be selected for study though the scientists are not necessarily cognizant of their practical significance."

Merton even goes so far as to suggest that the choice of research problems by scientists was "influenced" by a belief that "practical fruits" would result. Perhaps so, but the proposition is not clearly demonstrated. Great care is required to differentiate between what scientists say in justification of their work and what they actually do. From a Baconian viewpoint, of course, all true knowledge of nature must lead to better control of nature by man, so that man can seek power over his environment and his own condition through science.

Among the many insights in this section of the book, one in particular may be called to the reader's attention: Merton's discussion of "need." Acknowledging that certain requirements of invention in the military domain may indicate a response to the stimulus of the "needs" of war and noting the attention given to safety devices following the occurrence of accidents and disasters, Merton nevertheless insists that there is a multitude of human "needs" that "have gone unsatisfied throughout the ages." Indeed, the very word "need" is elliptical, since it seems to imply "realization or consciousness of need." Merton shows that an observer from one culture may often "detect a need in another society" simply because he comes "from a culture which has an established tradition of attempts to improve material welfare and to control nature"; this same "need," however, "*may not exist* for the members of the society under observation, precisely because of a difference in values and aims." This important insight helps to clarify the fact that some cultures and situations may *seem* to express a "need" that is not really a goal for the society in question insofar as the satisfaction of such a "need" by technological invention is not "part and parcel of the culture under consideration." Hence "need" is important as a source of invention only if it appears in a "cultural context... which places a high value upon innovation, which has a tradition of successful invention and which customarily meets such needs through technological invention rather than through other expedients."

Merton's primary example of "science, technology and economic development" is in the domain of transportation. This is a topic with immediate economic implications, since a "notable aspect of the economic expansion of this period was the need for more adequate means of transportation and communication." High on any list of scientific developments related to the technological "need" in this area is the problem of finding the longitude. There is no want of evidence that many prominent fellows of the Royal Society believed the improvement of navigation and of methods for determining the longitude was one of their major activities. Not only in England but also in all countries where science was practiced in the 17th century there was concern with the problems of the longitude. The list of the men of science interested in astronomy who concerned themselves with the longitude includes all the greatest men of the period, beginning with Galileo and going down through Wren, Hooke, Huygens, Hevelius, Mercator, Flamsteed, Halley, Leibniz and Newton. The careful studies instituted of the satellites of Jupiter (a notable feature of the activity of the Academy of Sciences in Paris) and of the moon's motion (by Flamsteed in practice and by Newton in theory) are testimony to the importance that the urgency of finding the longitude, notably at sea, had for the development of astronomy.

A major chapter of Merton's book deals with "science and military technique," exhibiting "the foci of scientific interests" that were "influenced" to an "appreciable degree" by the technological demands of war. Today we have become so accustomed to the close alliance between scientific research and military problems that Merton's documentation seems almost supererogatory. Yet in the 1930's this close connection did not exist to the same degree that it does today. Statistical evidence was required to support Merton's contention that military problems were of consequence, although he is bound to admit that many problems of military interest (such as the motion of projectiles and the impact of one body on another) "flowed directly from the intrinsic developments of science."

To study "the extent to which military, economic and technical influences were operative" Merton examined the minutes of the Royal Society for four years: 1661, 1662, 1686 and 1687. All topics discussed at these meetings were classified as "directly related" to "social or economic demands" or as "indirectly related" (when there seemed to be "some clear-cut connection with current practical needs, usually intimated in the context, but . . . not definitely so related by the investigators") or as "pure science." Less than half (41.3 percent) of the investigations reported during the four years "were devoted to pure science." The items that were only "indirectly related to practical needs" came to about 30 percent.

Merton addresses the question of the "influence" that economic (and social) factors may have on the "direction of scientific research." Of course, with few exceptions the activities of scientists were hardly motivated by "prospects of personal economic gain." Nor can one truly describe the scientists who met in the latter years of the 17th century at the Royal Society or in coffeehouses or the apartments of one of the fellows as "a group of 'economic men' jointly or severally seeking to improve their economic standing." Merton is perfectly aware that the factors influencing actual scientific research may be subtle and indirect.

Merton insists that one must not forget that the Royal Society's active members were essentially "a band of curious students coöperatively delving into the arcana of nature." Nevertheless, at their meetings they did discuss "without end technical problems of immediate concern for the profit of the realm." No doubt scientists were then aware of the possibility of "social acclaim" for "discoveries which promised profitable application" and which thus would be "heralded far beyond the immediate circle of virtuosi." There is much evidence of the ridiculing of all pure science that was removed from utility, such as weighing the air, whereas applause was given to scientific researches of more obvious practical potentialities, as in the areas of hydrodynamics, ship design, navigation and gunnery.

Merton finds support for his doctrine of the "appreciable directive influence of practical problems" in the fact that so many "scientists themselves point to the practical implications of their work." According to the principles of Francis Bacon, however, and according to a major view of science from the 17th century until now, any work in science that leads to an understanding of nature must eventually (or very likely will eventually) lead to a control of nature. There is no doubt that Newton contributed as much to the determination of the longitude by his studies of gravitation applied to the motion of the moon as did Huygens and Hooke, who applied scientific principles to the perfection of clocks. Whatever Newton may have said later on, there is absolutely no evidence in his early writings that he was thinking about problems of the longitude or that he had any primary motivation other than understanding the cosmic system and the forces that hold it together and produce the observed phenomena. Merton, however, believes that even "that 'purest' of disciplines, mathematics, held little interest for Newton save as it was designed for application to physical problems." This kind of generalization is not substantiated by the evidence now being made available in D. T. Whiteside's edition of Newton's mathematical papers.

The conclusion that Newton's pure mathematics is not so pure after all, since it is a mathematics designed for solutions of problems in physics, is a stage in the general thesis of utility developed by Merton: that physics itself is largely motivated by the technological demands of society. Even though Merton advances this thesis with caution and restraint and with many a caveat, he does conclude: "In general, then, it may be said that the contemporary scientists, ranging from the indefatigable virtuoso PETTY to the nonpareil NEWTON, definitely focused their attention upon technical tasks made prominent by problems of navigation and upon derivative scientific research. The latter category, however, needs careful delimitation. While it is true that a congeries of scientific investigation may be traced to technical demands, it is equally evident that much of this research can be understood as a logical development of foregoing scientific advance. It is only because the scientists themselves point to the practical implications of their work that one becomes inclined to accept the appreciable directive influence of practical problems."

This presentation gives a false picture of Newton and other scientists of that age. Like many others, Newton from time to time expressed satisfaction that his research might lead to useful results. Even then, however, he might have been echoing someone else's sentiments. A famous example may be found in a Scholium to Prop. XXIV, Book II, of Newton's *Principia,* on the mathematical problem of finding the "solid of least resistance" according to certain conditions of motion in a fluid. Newton comments that his result might "be of use in the building of ships." Cited by Merton and others, this conclusion seems not to have been original with Newton but to have been suggested to him by the Scottish mathematician John Craig. Far from being central to Newton's research, or in any way a motivating factor, this opinion was an afterthought. It was Craig, and not Newton, who envisioned the usefulness of the mathematical solution of the solid of least resistance, and Craig made a note to that effect in his own copy of the *Principia:* "The occasion for this afterthought I myself provided, when in Cambridge I proposed to the celebrated Author the problem of finding the most suitable outline for ships." This example well illustrates the fundamental problem of making precise the motivation of any scientist, particularly in terms of chance expressions concerning utility. Newton said on another occasion: "When I wrote my Treatise about our System, I had an Eye upon such Principles as might work with considering Men, for the Belief of a Deity, and nothing can rejoice me more than to find it useful for that Purpose."

Newton was indeed religiously motivated, as witness his extensive writings on theological and biblical subjects. He was a passionate student of alchemy and esoteric Hermetic philosophy. Not only was he driven by the creative force of his own personality and the logic of his own research into mathematics and physics but also he was obsessed by the view that he was making contact with a semisecret ancient tradition of learning; he even believed his own most original discoveries were rediscoveries of elements of a priscan wisdom. Furthermore, as Alexandre Koyré has noted, Newton's scientific research was both a conscious and an unconscious response to some basic philosophical problems of his age in relation to time, space, matter, movement and eternity. We have adequate testimony that Newton's interest in astronomy was aroused not by utilitarian considerations such as navigation but by astrology and by his desire to "test" this ancient pseudoscience.

Merton's analysis shows how the background of the age can be related to a Newton, but it fails to give any insight into the special qualities of personality and scientific genius that made him what he was. Yet Merton's insight into the role of Puritanism (or the "Puritan ethic") as a contributing factor, however inadvertent, "to the legitimacy of science as an emerging social institution" is of great consequence, even for understanding a Newton.

Newton's scientific career also illustrates an important observation by Merton on the significance of the formation of a scientific community. It seems likely that Newton would not have written the *Principia*, or at least would not have written it until much later, if there had not been a discussion by the London virtuosi (of the Royal Society) of the forces that produce Keplerian motion in the planets, followed by a visit by Halley to Newton to explore the topic. Newton was stimulated by Halley to set down his first researches on celestial dynamics. He was then encouraged by the Royal Society to complete these researches and to write them up in full for publication in a treatise to be issued under the sponsorship of the Royal Society.

The example of Newton shows that Merton's analysis, like all past and present sociological studies of science, may reveal qualities of the scientific enterprise as a whole but fail to illuminate the particular career of a single individual. Nonetheless, Merton has conclusively demonstrated that the growth of science in 17th-century England was not due to the chance concentration of a handful of geniuses but was conditioned by social and economic forces and a religious ethos that produced a "cultural soil... peculiarly fertile for the growth and spread of science."

Scientists, historians and sociologists still face questions about the ways in which social and economic considerations should enter into the analysis of the history of science. Merton's book, notwithstanding its panoply of graphs and quantitative data, does not yield wholly acceptable answers to today's inquirer. In particular, Merton does not give us any clue to the ways in which the psychological reaction of each scientist to the economic, religious and social values of his age may or may not determine the nature and direction of his scientific research. Even agreeing with the conclusions of Merton's investigations, one can assert that the greatness of English science and the glory of the Royal Society are not to be found in statistics but in the intellectual contributions of a handful of men of the highest genius: Newton, Hooke, Wren, Wallis, Halley, Ray and a few others. In that epoch science had not yet found our modern way of team or group research, in which first-class scientific work can be produced by men of less than first-class genius.

Shorter Reviews
by Philip Morrison

THE MAMMALIAN CELL AS A MICROORGANISM: GENETIC AND BIOCHEMICAL STUDIES *IN VITRO*, by Theodore T. Puck. Holden-Day, Inc. ($14.95). Bacterial cells have been the classical material for working out the modern molecular biology of the double helix and its meaning. Individual cells are deposited on nutrient jelly, where each survivor becomes the ancestor of a distinct clone of progeny, all genetically the same barring subsequent mutations, and so numerous that after a few hours the little patch of microscopic cells is visible on inspection with the unaided eye. That colony can be sampled and the sibling cells treated differently—by regrowth on differing nutrients, by radiation and so on. Genetic experiments with fruit flies, let alone man, take a week or two and use 10,000 organisms, each of which must be examined, say, for eye color. Bacterial genetics, however, deals with billions of organisms that grow overnight; rare biochemical events are thus easily observed. This thin but rather detailed volume explains how over the past 15 years or so the large and much more delicate cells of mammals, including those of our own species, have been made available for study in ways strictly parallel to the genetic study of the hardy, minute bacteria.

The technique did not come easy. At first tissues grew well, but isolated cells simply would not grow. A brilliant experiment of 1948 showed that cells made some substance that helped their neighbors to grow. Step by step methods were found to allow routine plating of single cells into Petri dishes like so many bacteria, there to grow into colonies with a yield of up to 80 percent. The starting cells can be taken from skin or blood or "a variety of normal human and animal tissues." The culture will grow actively for "approximately 100 generations" before slowing down or stopping altogether (cellular old age?). The trick is to avoid all trauma, maintaining temperature, humidity and the like with great care. "It is difficult for a person familiar only with the relatively rugged growth of... microorganisms to appreciate the enormous sensitivity of mammalian cells when removed from their natural habitat," the result of a few hundred million years of adaptation.

The nutrient medium for reliable growth is all but completely chemically defined. For the much used hamster ovary cell, say, it contains known simple molecules: 16 amino acids, seven vitamins, 10 inorganic salts, glucose and another sugar derivative, linoleic acid and only one polymeric derivative of mammalian blood serum—a protein-carbohydrate compound. This alone spoils the elegance of the medium, and yet it seems essential (although it may well turn out that only some detachable portion of the big molecule, which is built of a couple of hundred peptide and sugar blocks, is truly necessary). Cells in a culture seem to secrete a related compound, and so some cells can now grow in a medium free of any proteins. The cells divide in periods of from 10 to a few tens of hours, always best near blood heat. Cells of many types will grow in layers lying on top of one another and not just directly on the nutrient surface, although this property is not shared by all kinds of cells. "It is intriguing that many cells which normally grow" only in monolayers began to pile up "after treatment with carcinogenic agents."

Human cells are large and complex compared with bacteria. The genetic material in man is densely folded into chromosomes a few microns long, holding in all some six feet of DNA molecular tape; bacteria have instructions a couple of millimeters long. Studies of chromosomes are perhaps the clearest and most successful of all the results of the cell-culture technique to date. The chromosome count in man, held to be 48 from 1926 to 1956, has now been verified in a large number of skin-cell cultures, confirming "with overwhelming clarity" that the newer number of 46, first found in a few short-term cultures from fetal material, was correct. Nowadays the count can be taken clinically, using blood cells, within 48 hours.

Those complex fibers of DNA, intricately cemented by proteins, that form the human chromosomes are fateful. Some 40 specific disease-producing anomalies are now well documented, and many are listed in the text. One percent of the male population of institutions for the mentally deficient is accounted for by persons who possess both a male and a female sex-chromosome complement. Lacking an end portion of one chromosome of pair 22, a patient will almost certainly develop a chronic form of leukemia. One chromosome or three always mean pathology; the simple lack of one copy of a single gene often goes unnoticed, but an urgent need for chromosomal balance is clearly present. The all-or-none idea of gene information is inadequate here: many genes must regulate development quantitatively. In living people we see anomalies almost only in seven pairs of chromosomes out of the entire set; visible defects in the rest lead to spontaneous abortion.

There is a good account of the remarkable properties of the human sex chromosomes. Female tissues all yield cells with two *X* chromosomes, of course, but only one of these chromosomes is active; the other condenses into a strongly staining region, the "Barr body." It was shown 10 years ago that single cells from skin samples of women who had only one gene on the *X* chromosome for a particular enzyme (judged by their having borne several male children without the enzyme) gave rise to clones of which half possessed the enzyme and half lacked it. Every woman is thus a mosaic of clones, in each of which only one of the two *X* chromosomes functions. The random decision between *X*'s is made early in embryonic life by each embryo cell sometime after the embryo is multicellular. In this way women retain balance and yet avoid the sex-linked genes that cause much disease by mutation in males. Otherwise all cells of the normal body probably contain full information for development, possibly made latent in one way or another. (There is mention of true differentiation to the adult form from a skin cell of a frog once the nucleus is put into an enucleate egg. The external culture of mammalian sex cells, which can differentiate on reimplantation, is not discussed.) A complete map of the human genes is one goal for these new genetic methods.

Radiation damage has been studied most revealingly in cell culture. After moderate radiation many cells grow to giant size, 10 times normal or more, often becoming easily visible to the unaided eye. They make no clones; they can grow well but they cannot divide. Two abnormal irradiated cells are shown, looking like two fish hooked to the ends of a short line; a chromosomal bridge stubbornly unites the pair, which normally could separate to consummate orderly division. Bacterial cells are far more resistant to radiation; our cells present a large fibrous chromosomal machinery to the stream of damaging ions.

This very interesting book is written as a textbook. Its level of detail is not at the requisite fullness for actually growing the cultures or doing the experiments. It is aimed at graduate students in biology, but one would expect well-read undergraduate students, or others with an interest in biology, to profit from it. Much of the appendix apparatus, a wide but brief review of molecular biology, seems rather oversimplified for any audience that might hope to grasp most of the text. Professor Puck has concernedly provided each chapter with an appraisal of the "human implications" of his work, whose entire burden plainly implies startling outcomes not very far ahead, which we will need to foresee and to embed within a "new intellectual and moral synthesis," to cite his closing editorial.

THE OCCULT SCIENCES IN THE RENAISSANCE: A STUDY IN INTELLECTUAL PATTERNS, by Wayne Shumaker. University of California Press ($15). Giovanni Baptista della Porta published his Latin runaway best seller, *Twenty Books of Natural Magic,* when Galileo was a young lecturer at Pisa. The book ran through 27 editions and was almost immediately translated into Italian, French, Dutch, Spanish "and apparently Arabic." We know it best for its account of the use of lenses; it was almost the first book to refer to them. Lenses had been widely used in spectacles for nearly 300 years but were entirely ignored by the learned. Della Porta even challenged the scholars; he remarked that no one had explained either the effects of lenses or the reasons for them. It has been argued by Vasco Ronchi, the distinguished Florentine physicist and historian of optics, that it was Della Porta's book that led to the first telescope—not Galileo's, of course, but one made in 1590 by an unknown Italian. That instrument was seen in 1604 by the Dutch spectacle grinders, who then produced the toy telescopes whose potential inspired Galileo to develop better ones and turn them to the Tuscan sky.

It was an extraordinary chain of natural magic! Della Porta's treatise itself, however, is almost without readability or credibility today. It is decent enough, the whitest of magic; "no daemonic powers whatever are solicited." All it tells depends on relations and forces seen as being objectively present in nature; there are no prayers, no invocations, no ceremonies. The Neapolitan author searched the world of "Libraries, Learned Men, and Artificers" to bring back "a survey of the whole course of Nature," so that the magician, "a very perfect Philosopher," might "do strange works, such as the vulgar sort call miracles." The magician should be versed in mathematics and astrology and understand optics, minerals, herbs and medicine. Finally, he "must also be rich." Then the seeming miracles can flow. These are passing strange, although supported from analogy or old and traditional wisdom. The tongue of the chattering goose placed in a woman's bed will make her "utter her night-secrecies." Wild olive is an antidote to a chameleon an elephant has eaten by accident. Dromedary bred with boar yields the two-humped camel. And so on, for 20 brief "books," with here and there the nugget of a real lodestone or a burning glass, or even a plausible herbal remedy.

It is plain that the thought of the Renaissance, the centuries in which the modern mind was brought to term, was permeated with such occult material. It is not intrinsic strangeness alone that causes the recipes to lose their hold on us; the story of the telescope and of its fruit for the mind is surely as implausible as curing a mad dog's bite by "a Wolves skin put upon any one that is bitten." After all, the cure we use now—and it very often works—is to mingle with the bloodstream of "any one that is bitten" the ground spinal cord of a mad rabbit, treated with stinking formaldehyde. The difference is a deeper theoretical structure and above all a nexus of trials and tests, much more intricate than the citations from a few ancient writers or the single forced analogy on which the old authors rested their case.

"How did the remotest ancients come to know so much?" It is not just scripture that forms the basis for esoteric knowledge; it is as likely to be some Latin poet, palpably writing fiction. The senses were not seen as reliable guides; illusions are ever their lot. Yet the past, somehow closer to divine inspiration, is a true guide; even its romances are veiled truth. "The real ground for belief was faith . . . in authorities."

So does Wayne Shumaker, professor of English at Berkeley, critic and literary analyst, open to us the tangled growths of the occult. He treats five such sciences: white magic, witchcraft, astrology, alchemy and the curious intellectual movement around certain Greek and Latin manuscripts that trace back to an Alexandrian school of mysticism but were ascribed to one mysterious master, Hermes Trismegistus. There is, of course, an embarrassment of many books on each of these topics, except perhaps the last. But here one sees the occult systems as a whole, not minutely, for the most part, but as theoretical and practical structures with their own aims and methods, to be sifted for meaning and coherence. Often Shumaker compares the doctrines of contending authors; it is not to be thought that all minds agreed on the doctrines or on their validity. He provides long citations, and he has brought his reader indispensable gifts: his expertise and his patience. He has read the dusty old books, in the tongues of all Europe or in Latin, and he has set down what he finds there with many long citations.

He is no believer. Indeed, he is endearingly candid about his concern that "young people have revolted strongly against reason." Therefore he offers a very clear refutation of astrology, for example, on internal and on empirical grounds (much of his stance is taken from the contemporaries of his astrological sources), along with his clear, explicit and compact account of a horoscope. "A decade ago I would not have written [such refutations], and in another decade I hope they will again become unnecessary. At the moment, I could not in conscience omit them."

The long-festering barbarism of the treatment of witches is particularly painful. It was not the Dark Ages but the High Renaissance that tortured so many poor, uneducated, outcast and half-crazy old women. A couple of hundred thousand people, mostly women, were burned between the time of the voyages of Henry the Navigator, when Pope Innocent VIII called officially for the extirpation of witchcraft, and "the dying down of fanaticism" in about 1700, after the frenzy had passed even in distant Massachusetts. The scholars' theories justified it all; "each used his sources exactly as his modern counterpart does, balancing one set against another or supporting his own views by multiple citations."

Still, there are crucial differences. We cite ancient authority more as evidence of ancient belief and less as source of fact. Present experience was also then regarded as being important, but it was seldom treated critically. Any statements were taken as being true; old wives' tales were the main burden of evidence, the more acceptable because they resembled the reports of the distant past. There were skeptical and compassionate writers too, and in the end they prevailed. "Ultimately, except in cultural backwaters, the battle for sanity was won." Yet great Martin Luther and lucid Sir Thomas Browne accepted the existence of witches. Montaigne (from a 1603 translation) said what is in the modern mind: "How much more naturall and more likely doe I finde it, that two men should lie, than one in twelve hours [should] passe with the windes, from East to West? . . . It is an over-valuing of ones conjectures, by them to cause a man to be burned alive."

The study is a library project; the laboratory aspect of alchemy, say, is lightly treated. But the annotated lists of references, the period illustrations, the recognition that serious and able minds once differed greatly from our own and the absence of contaminating traces either of condescension or of credulousness give this absorbing volume a special authority and a place on the shelves of any reader or any library where the history of modern thought is relevant. This is a time when the failures of reason lead many who should know better to seek refuge not in clarity but in unreason; that no comfort lies therein can be learned from this honest history.

THE SCANNING ELECTRON MICROSCOPE: PART I, THE INSTRUMENT, by C. W. Oatley. Cambridge University Press ($16.50). THE SCANNING ELECTRON MICROSCOPE: WORLD OF THE INFINITELY SMALL, by C. P. Gilmore. New York Graphic Society ($15.95). Aim a narrow flashlight beam ahead of you into a pitch-dark room. Scan the room systematically, recording at each position of the spot the total amount of light; never mind where it comes back from. That is a close analogy to the scanning electron microscope. Here, however, the specimen is held in good vacuum, the spot is an electron beam with 20 kilovolts of energy, perhaps 100 angstroms in diameter, and the record is a display on a 1,000-line television screen completely scanned each second.

The results are extraordinary. Since the specimen is seen by reflection or secondary processes of some kind, no thin section is needed. Since only the near-parallel exploring beam uses the image-forming properties of magnetic lenses, the depth of focus is great, much larger than that of any optical microscope of comparable magnification. Magnification is easily varied merely by changing the scan track. It is practical to vary the magnification smoothly and quickly from perhaps 20 diameters to 50,000 without refocusing. (The scanning coils of beam and display tube are effectively connected in series in order to maintain synchronism.)

The results are by now ubiquitous, although not commonplace. They allow the eye to roam the small world in depth, freed from those puzzling thin layers that were formerly the hallmark of high magnification. How it is that this scanning process produces such readable pictures, how we recognize at once the topography of "hills, depressions, crevasses, holes and edges" is by no means evident. The electrons reach the specimen from one sharply defined direction, which provides our viewpoint. The narrow beam reaches deep into holes and crevasses, so that their walls shine out to us even if electrons must come back by roundabout and diffuse pathways. The inclination of the surface does not directly affect the image, because the effects of beam and angle and image-foreshortening cancel, but whenever the beam angles in, more secondary electrons can reach the surface to be recorded. "The formation of contrast is a most complex business."

Professor Oatley has been working on the scanning microscope at Cambridge since 1948. The root idea goes back to German work more than a decade earlier, but the development of practical devices is a long story of understanding and improvement, the rise of signal and the decline of noise. A narrow probing beam cannot carry much current, because the demagnifying lenses that form it cannot in principle compact electrons both in position and in direction, so that practical limits are set by the random sideways motion of the electrons boiling off the cathode. A tungsten hairpin running white-hot in a triode gun is still the standard cathode, although cold emission from a sharp point also has merit. One worker has used lanthanum hexaboride rod as his electron source; its lower working temperature means less thermal motion and a tenfold increase in brightness. Detectors vary too. The standard is the scintillating plastic viewed by a photomultiplier, but semiconductors are used, and simple amplification of the current to the entire specimen is also useful. The usual beam current amounts to only about 10 million electrons per second.

The book is a self-contained introduction to the physics of the instrument and its construction principles. The design limitations are made clear and quantitative and the major components are discussed in some detail, with the reasons for design choices given. The aim is to promote the intelligent use of the device, making the basic ideas as simple as possible so that a user can know what lies behind the many variables under his control. But the proof of the microscope is in the pictures from it, and Professor Oatley gives us none; he is a modest parent.

The second volume, issued by a publisher whose main output is handsome books of reproductions of works of art, remedies the lack. A number of books have already appeared—some reviewed in these columns—that display fine sets of scanning electron micrographs. They were all restricted in scope, usually the work of microscopists devoted to some discipline. The compiler of the present volume is an experienced editor and writer of science news for the general reader. He has brought together from laboratories in many countries a catholic collection of close to 200 excellent halftones of micrographs, many of them a full page in size and some even two-page spreads. They display the diversity of the world at magnifications of from 30 diameters to 52,200, all with the convincing illusion of depth that this wonderful instrument generates. Here are a red blood cell netted in a tangle of fibrin; a clumsy-looking needle eye through which the elegant harpoon end of a bee's stinger has been threaded; a single human hair, round and scaly as a palm tree; a woolly-looking chromosome; a landscape of picket fences made of tungsten fibers growing out of the grain boundaries of a uranium dioxide matrix; a mite, ubiquitous in house dust yet all but unseen; a burred pollen grain of morning glory; the rolling ball of a common pen. Not for three centuries now has it been so agreeable for sightseers in the minute world, here displayed not merely in the ultrafine detail shown by the transmission electron microscope but at a scale that reveals sculptural forms at the micron level or just below it. The grooves of a phonograph record bear the individual strokes of the cutting point, the salt cubes have texture, the paramecium is furry.

INDEX OF ADVERTISERS

FEBRUARY 1973

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

RECONNAISSANCE AND ARMS CONTROL

OBSERVATION SATELLITES: PROBLEMS AND PROSPECTS. Amrom H. Katz in *Astronautics;* April, June, July, August, September, October, 1960.

SECRET SENTRIES IN SPACE. Philip J. Klass. Random House, Inc., 1971.

RECONNAISSANCE, SURVEILLANCE AND ARMS CONTROL. Ted Greenwood. Adelphi Paper Number 88, The International Institute for Strategic Studies, 1972.

THE ORIGINS OF NERVE-CELL SPECIFICITY

CHEMOAFFINITY IN THE ORDERLY GROWTH OF NERVE FIBER PATTERNS AND CONNECTIONS. R. W. Sperry in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 50, No. 4, pages 703–710; October, 1963.

DEVELOPMENTAL NEUROBIOLOGY. Marcus Jacobson. Holt, Rinehart and Winston, Inc., 1970.

DEVELOPMENT AND STABILITY OF POSITIONAL INFORMATION IN *XENOPUS* RETINAL GANGLION CELLS. R. K. Hunt and Marcus Jacobson in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 69, No. 4, pages 780–783; April, 1972.

FISHES WITH WARM BODIES

"THE WONDERFUL NET." P. F. Scholander in *Scientific American,* Vol. 196, No. 4, pages 96–107; April, 1957.

WARM-BODIED FISH. Francis G. Carey, John M. Teal, John W. Kanwisher, Kenneth D. Lawson and James S. Beckett in *American Zoologist,* Vol. 11, pages 137–145; 1971.

THE CHINESE LANGUAGE

A GRAMMAR OF SPOKEN CHINESE. Yuen Ren Chao. University of California Press, 1968.

CLIBOC: CHINESE LINGUISTICS BIBLIOGRAPHY ON COMPUTER. William S-Y. Wang and Anatole Lyovin. Cambridge University Press, 1970.

THE COMPUTER IN CHINESE LINGUISTIC RESEARCH. William S-Y. Wang, Stephen W. Chan and Benjamin K. Tsou in *Proceedings of the USA-Japan Computer Conference,* 1972.

DOC, 1971: A CHINESE DIALECT DICTIONARY ON COMPUTER. Mary L. Streeter in *Computers and the Humanities,* Vol. 6, No. 5, pages 259–270; May, 1972.

A SYNCHRONIC PHONOLOGY OF MANDARIN CHINESE. Chin-Chuan Cheng. Mouton Publishers, 1973.

THE MICROSTRUCTURE OF THE OCEAN

AN ANALYSIS OF THE STIRRING AND MIXING PROCESSES IN INCOMPRESSIBLE FLUIDS. Carl Eckart in *Journal of Marine Research,* Vol. 7, No. 3, pages 265–275; 1948.

A NEW CASE OF CONVECTION IN THE PRESENCE OF COMBINED VERTICAL SALINITY AND TEMPERATURE GRADIENTS. J. S. Turner and Henry Stommel in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 52, No. 1, pages 49–53; July, 1964.

WAVE-INDUCED SHEAR INSTABILITY IN THE SUMMER THERMOCLINE. J. D. Woods in *Journal of Fluid Mechanics,* Vol. 32, Part 4, pages 791–800; June 18, 1968.

THE VERTICAL MICROSTRUCTURE OF TEMPERATURE AND SALINITY. M. C. Gregg and C. S. Cox in *Deep-Sea Research and Oceanographic Abstracts,* Vol. 19, No. 5, pages 355–376; May, 1972.

OCEANIC FINE STRUCTURE. Thomas R. Osborn and Charles S. Cox in *Geophysical Fluid Dynamics,* Vol. 3, No. 4, pages 321–345; July, 1972.

THE CRASHWORTHINESS OF AUTOMOBILES

BASIC RESEARCH IN AUTOMOBILE CRASHWORTHINESS—SUMMARY REPORT. P. M. Miller and R. P. Mayor. Calspan Corporation Report No. YB-2684-V-6, November, 1969.

1970 INTERNATIONAL AUTOMOBILE SAFETY CONFERENCE COMPENDIUM. Society of Automotive Engineers, Inc., 1970.

ACCIDENT FACTS. National Safety Council, 1972.

EXPERIMENTAL SAFETY VEHICLES: WHERE DO WE GO FROM HERE? *The SAE Journal of Automotive Engineering,* Vol. 80, No. 8, pages 19–27; August, 1972.

NEW PASSIVE RESTRAINT SYSTEMS. *The SAE Journal of Automotive Engineering,* Vol. 80, No. 8, pages 30–39; August, 1972.

METAL-VAPOR LASERS

CW LASER ACTION ON 24 VISIBLE WAVELENGTHS IN SE II. W. T. Silfvast and M. B. Klein in *Applied Physics Letters,* Vol. 17, No. 9, pages 400–403; November 1, 1970.

NEUTRAL GAS LASERS, Colin S. Willett, IONIZED GAS LASERS, W. B. Bridges and A. N. Chester in *Handbook of Lasers.* The Chemical Rubber Company, 1971.

NEW CW LASER TRANSITIONS IN SE II. M. B. Klein and W. T. Silfvast in *Applied Physics Letters,* Vol. 18, No. 11, pages 482–485; June 1, 1971.

PENNING IONIZATION IN A HE-CD DC DISCHARGE. W. T. Silfvast in *Physical Review Letters,* Vol. 27, No. 22, pages 1489–1492; November 29, 1971.

ROTATION IN HIGH-ENERGY ASTROPHYSICS

ON THE NATURE OF PULSARS: I, THEORY. J. P. Ostriker and J. E. Gunn in *The Astrophysical Journal,* Vol. 157, No. 3, Part 1, pages 1395–1417; September, 1969.

THE CRAB NEBULA: PROCEEDINGS OF THE 46TH INTERNATIONAL ASTRONOMICAL UNION SYMPOSIUM. Edited by R. D. Davies and F. G. Smith. Springer-Verlag, Inc., 1971.

PULSARS: STRUCTURE AND DYNAMICS. M. Ruderman in *Annual Review of Astronomy and Astrophysics,* Vol. 10, pages 427–476; 1972.

MATHEMATICAL GAMES

THE GAMOW-STERN ELEVATOR PROBLEM. Donald E. Knuth in *The Journal of Recreational Mathematics,* Vol. 2, No. 3, pages 131–137; July, 1969.

DIALOGUE ABOUT PUZZLES. Kobon Fujimura and Michio Matsuda. Tokyo, 1971.

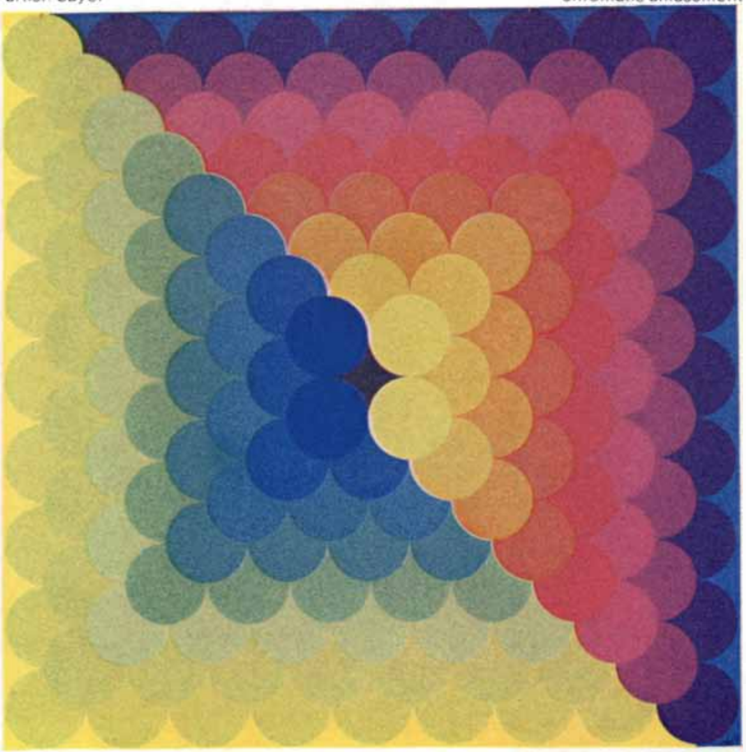
THE ART OF COMPUTER PROGRAMMING: VOL. III, SORTING AND SEARCHING. Donald E. Knuth. Addison-Wesley Publishing Co., Inc., 1973.

THE AMATEUR SCIENTIST

THE ARCHITECTURE OF MOLECULES. Linus Pauling and Roger Hayward. W. H. Freeman and Company Publishers, 1964.

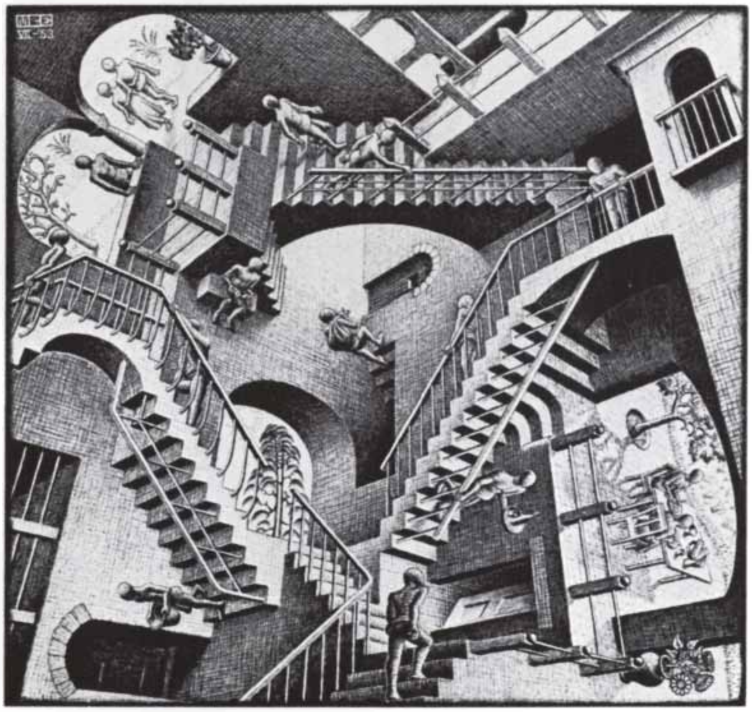
BIOCHEMISTRY: THE MOLECULAR BASIS OF CELL STRUCTURE AND FUNCTION. Albert L. Lehninger. Worth Publishers, Inc., 1970.

artist: bayer chromatic amassment



The ideal
A society in balance. A healthy, well housed, fully employed peacetime population—with clear air, clean water and equal opportunity for everyone.

artist: escher



The real
We move in different directions, disregarding our neighbor's goal. We dilute our efforts. We fail to reach the equilibrium our strength could give us.

Achieving national goals requires a balanced effort. We must continue to seek new ways to reduce air and water pollution . . . raise the standard of living of men and women whose potential contribution to society is not being realized . . . and maintain a sound economy, which will be necessary to achieve environmental and social goals.

Above all, we must broaden our perspective to weigh all our goals in making decisions. For these goals are interrelated. We cannot afford to pursue any one of them at the cost of another.

AtlanticRichfieldCompany