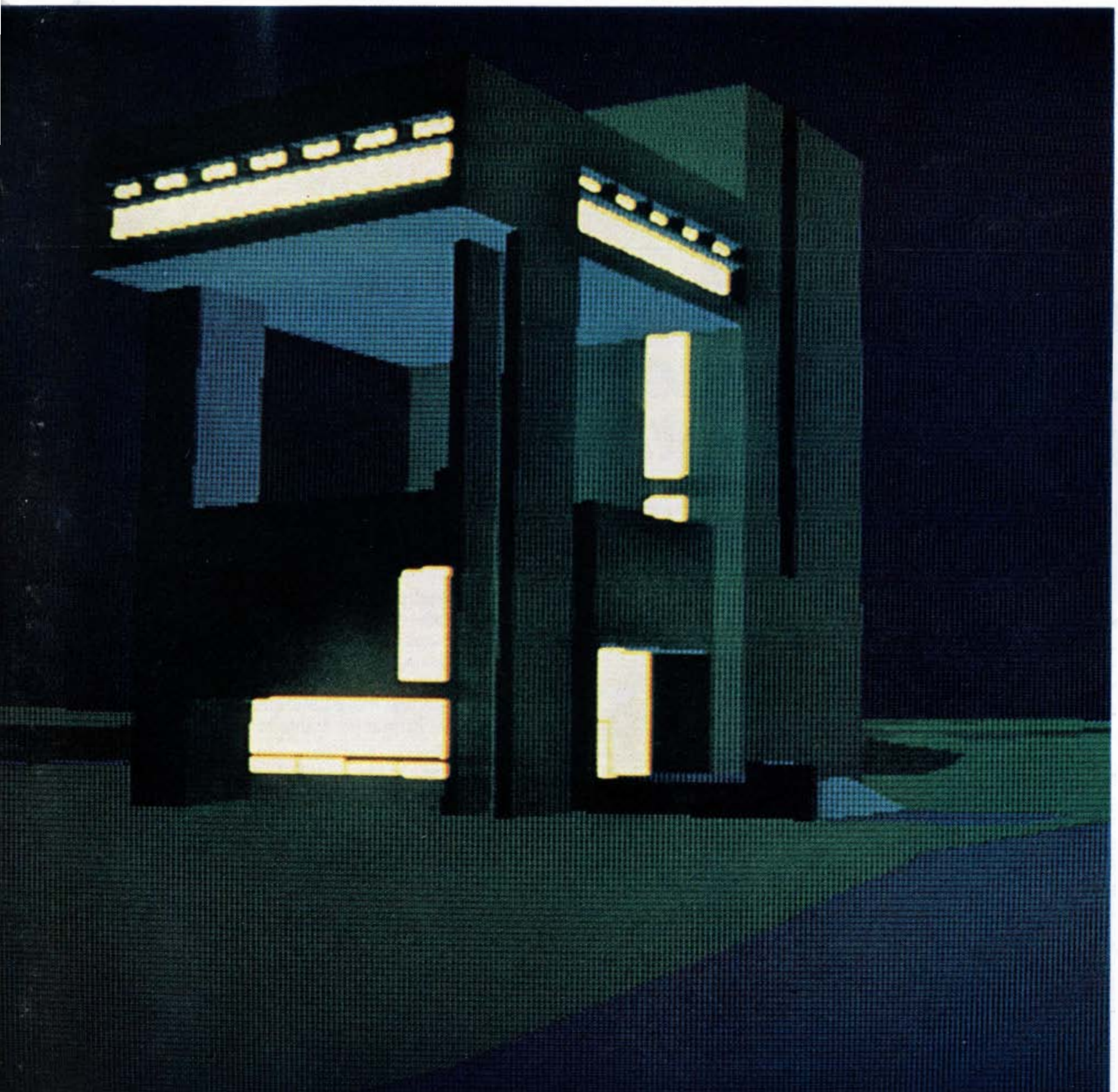
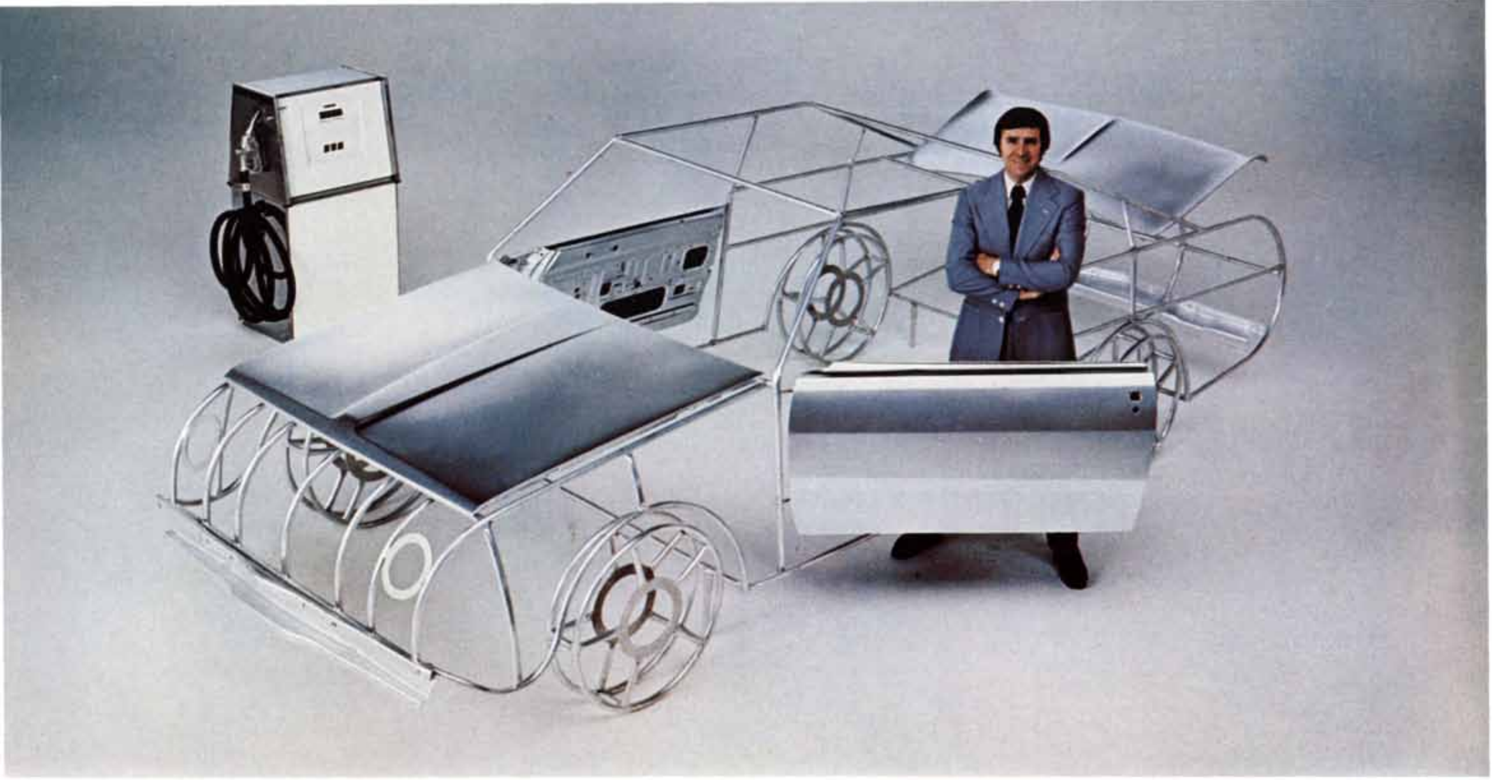
SCIENTIFIC AMERICAN



COMPUTER GRAPHICS IN ARCHITECTURE

ONE DOLLAR

May 1974



Soon, your car will be a lot lighter. On the road, and on your wallet.

American cars must become smaller and lighter and less expensive to operate. Part of their diet to lose weight will be strong, lightweight aluminum alloys.

Four aluminum parts alone could save 345 pounds. This lighter car would burn less gas. We would also save on brake wear, tire costs and even registration in some states.

Let's take a 3,600-pound automobile and put it on an aluminum diet. We'll change the hood, trunk, doors and bumper reinforcements from steel to aluminum. In steel they weigh 380 pounds. In aluminum, only 150 pounds. That saves us 230 pounds.

There's more. We could save an additional 115 pounds through lighter supporting components that wouldn't have to support so much weight in the first place. That brings our total weight savings to 345 pounds. And pares the car down to a lightweight 3,255 pounds. If the car had four doors, aluminum could save an additional 105 pounds.

Even after the car has served us and gone on to scrap, we could use it. Aluminum brings about $200 a ton as scrap, so it's worth recycling. And when more of it goes into our cars, they become more valuable as scrap. An additional incentive to recycle used-up automobiles.

If you'd like a more in-depth look at how aluminum in automobiles helps in the conservation of energy, please write for our brochure, *Energy, Aluminum and the Automobile.* We're Aluminum Company of America, 352-E Alcoa Building, Pittsburgh, Pa. 15219.

The reasons for using aluminum are found in aluminum itself.

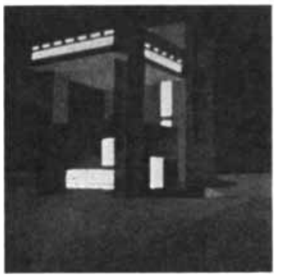
ALCOA

Established 1845 SCIENTIFIC AMERICAN May 1974 Volume 230 Number 5

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $12 PER YEAR, U.S., POSSESSIONS AND CANADA; $15 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O PUBLICATIONS DISTRIBUTION SERVICE, ST. KATELIJNEVEST 14, B2000 ANTWERP, BELGIË.



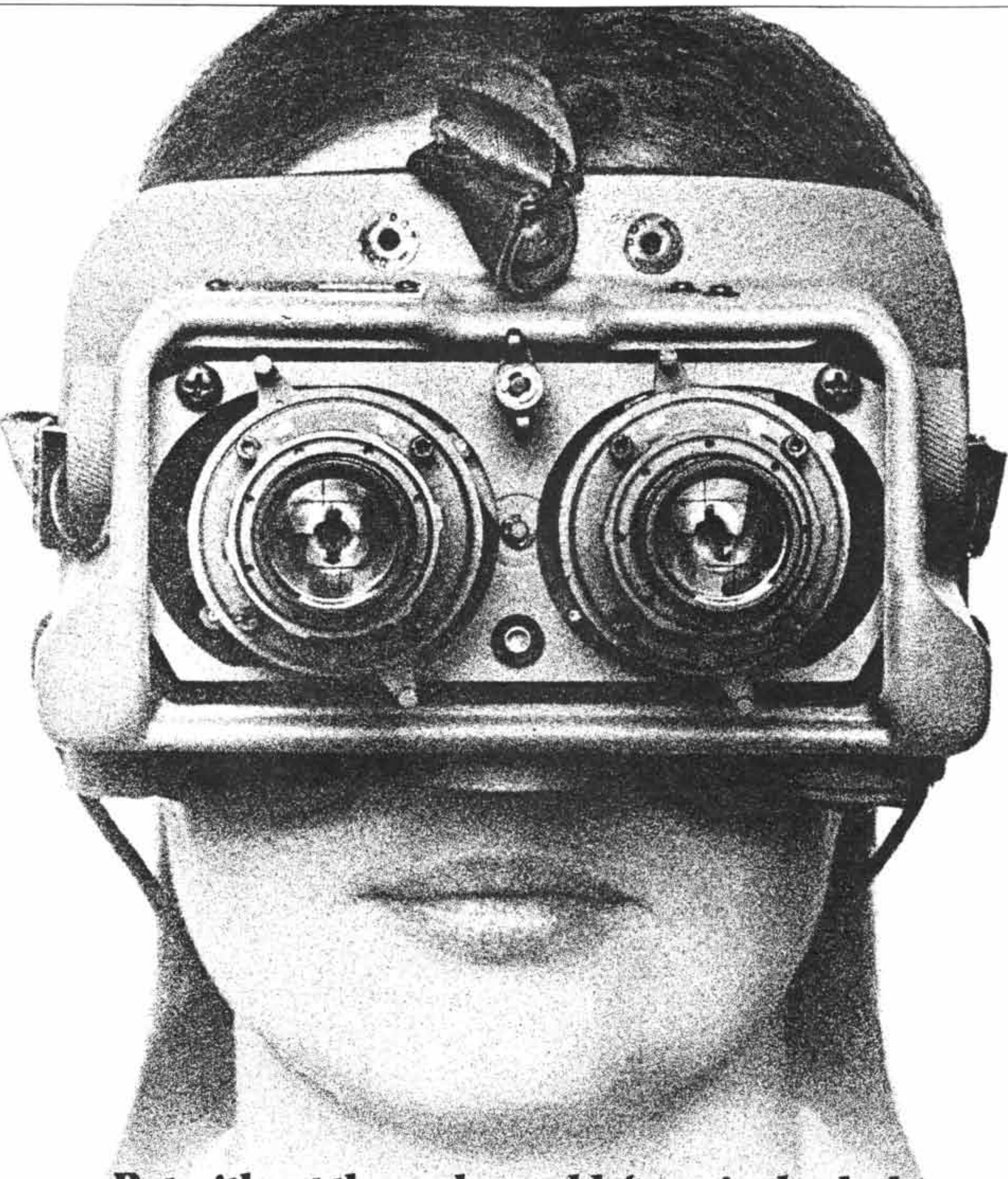
THE COVER

The illustration on the cover is a computer-generated architectural display showing the finished appearance of a university building that, at the time the display was made, had not been completed (see "Computer Graphics in Architecture," page 98). The building is the Herbert F. Johnson Art Museum at Cornell University. Under a program developed by Cornell and the General Electric Company the computer can be programmed to generate views of such a building from various perspectives and to simulate how it would look from different vantage points as one approached it on foot. The displays, which can offer 64 variations of color, appear on a television screen. Computers can also generate working architectural drawings.

THE ILLUSTRATIONS

Cover photograph by Donald P. Greenberg, Cornell University

Page	Source
21–31	Ilil Arbel
32	Martin Biddle, Winchester Research Unit of the Winchester Excavations Committee
34–39	Eric Mose
40–42	Martin Biddle, Winchester Research Unit of the Winchester Excavations Committee
43	Eric Mose
45	National Zoological Park, Smithsonian Institution
46	Walter Chandoha
47	Bunji Tagawa
48	R. W. Guillery, University of Wisconsin Medical School (*top*); Bunji Tagawa (*bottom*)
49–54	Bunji Tagawa
63	George V. Kelvin
64	Ben Rose
65–77	George V. Kelvin
79–86	Dan Todd
89	G. H. Mills, National Engineering Laboratory, Glasgow
90	R. W. Wright, Trans-Sonics, Inc.
91	Enid Kotschnig
92	Enid Kotschnig (*top*); Trans-Sonics, Inc. (*middle*); William A. Roth, Coates and Welter Instrumentation Corp. (*bottom*)
93	Harry Koba, Massachusetts Institute of Technology (*top*); Enid Kotschnig (*bottom*)
94	G. H. Mills, National Engineering Laboratory, Glasgow (*top*); Enid Kotschnig (*bottom*)
95	Enid Kotschnig (*top*); Arthur W. Ruff, Jr., National Bureau of Standards (*bottom*)
96	G. H. Mills, National Engineering Laboratory, Glasgow (*top*); Enid Kotschnig (*bottom*)
99–100	Donald P. Greenberg, Cornell University
101–102	Alan D. Iselin
103	Alan D. Iselin (*top*); Donald P. Greenberg, Cornell University (*bottom*)
104	Donald P. Greenberg, Cornell University
106	Saphier, Lerner, Schindler, Environetics, Inc.
109	Wolfhard Schlosser, Ruhr-Universität, Bochum, Germany
110–114	Allen Beechel
115	Lick Observatory
116–117	Allen Beechel
121–124	Ilil Arbel
127–130	Jerome Kuhl



Technology: Abandon,

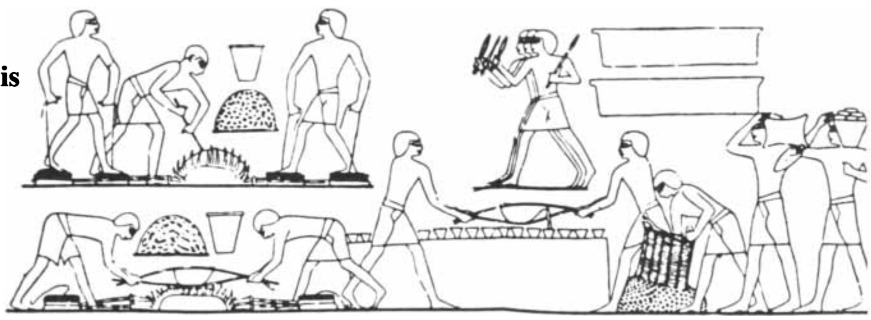
If you think we are living in an age of technology, you are perfectly right. If you think this is a phenomenon of recent centuries, you are entirely wrong.

Technology is simply the knack of doing things with objects that are not part of your body. If you try to crack a nut with your teeth, you are being natural. If you hit the nut with a rock, you are employing technology.

Man's realization that he could adapt his material environment to enhance his own comfort gave birth to technology.
The single most crucial step in man's history was his discovery that *tame* fire could furnish warmth and light. This use of fire represented the first absolute distinction between man and all other creatures. The technology allowed him to tailor the environment to his immediate needs.

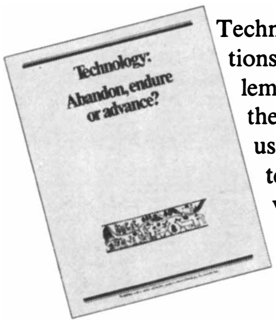
But fire, however efficiently and carefully used, produced undesirable side-effects which fouled the air. Many an early fire-user must have longed for the clean air of a tent or cave not polluted by smoke.

The choices were to abandon the use of fire, endure the pollution, or think of a way of making fire without pollution.
The last option is surely the most reasonable, and it was, indeed, the one taken. Some chimney-like device was developed which would allow the smoke to escape into the open air where it was diluted and carried off.



This is an example of how problems caused by technology are solved by additional technology. This, in turn, may produce new problems. Man's fires have, in fact, now increased in quantity and intensity to where the entire earth's atmosphere

We believe in the promise of technology.



Technology has provided the solutions to many of society's problems. It is also often blamed for them. Some people would have us believe that there's too much technology today — that this very abundance is the basis for so many of our ills.

We believe the solutions to problems lie in the intelligent application of technology—that a misunderstanding of technology's role in man's advancement can only lead to a diminution of his ability to provide solutions.

Therefore, we're sponsoring a series of "white papers" to foster a clearer understanding of technology. Excerpts from the first of these papers are published above. If you'd like a copy of the complete text about technology's role in man's development, please write Gould Inc. "Technology," Dept. S, 8550 W. Bryn Mawr Avenue, Chicago, Illinois 60631.

LETTERS

Sirs:

As J. D. North points out in his excellent article ["The Astrolabe," by J. D. North; SCIENTIFIC AMERICAN, January], the ancient mariners' astrolabe was "a crude device, serving chiefly to find the altitude of the sun, moon or stars." It may be of interest that a modern descendant of the more versatile original astrolabe is still very much in use in the U.S. Navy, but for computations rather than for observations.

Progress (sic) has replaced the beautiful brass and gold filigree work with sturdy plastic, but the ancestry of the design remains. The rete is a white plastic sheet, with 57 stars located and identified in black, and is reversible for observations in the Southern Hemisphere. The climates are a series of transparent disks overprinted with fine almucantars and azimuth lines. The rete is bordered with a scale for sidereal hour angle that is used to orient the climate.

In its commonest use the rete is annotated with current planetary positions (a modern addition). The appropriate climate is then selected and positioned for the sidereal hour angle of the next planned observation (for example an evening-star fix). The device is then consulted by the navigator to select the best stars to "shoot" that evening, based on their calculated altitude and azimuth. A second important use is on evenings with scattered clouds. Here the stars are first observed, then their identity is determined with the device; thus it is not necessary to see the entire constellation for identification.

This modern descendant of the astrolabe comes packed in a leatherette case with a rete, nine climates, a planet-plotting device and instructions. It is available from the Naval Oceanographic Office in Washington (and sells for about $6) as the "Star Finder and Identifier, Principle of Rude and Collins." The horse has been replaced by a plastic peg, and the instructions make no mention of the instrument's heritage, but although it is a bit degenerate, it remains an astrolabe.

LT. COMDR. R. S. PETERSON

U.S.S. *Iwo Jima*
At Sea

Sirs:

The simultaneous ending of the secular trend in age of menarche and the secular trend in body size, reported by Professor Tanner ["Growing Up," by J. M. Tanner; SCIENTIFIC AMERICAN, September, 1973] for economically well-off populations, may be more than a coincidence. The simultaneous cessation of both trends would be expected, because it has been shown that the weight dependency of sexual maturation applies to the human female, as well as to rats, mice, pigs, cattle and other domestic animals, which also show a secular trend in age of coming into estrus, or first heat.

Professor Tanner's explanation that better nutrition is most likely responsible for the secular trend in age of menarche follows from the weight-dependency relationship. Charles Darwin noted the relation of early maturation and fertility to abundant food supplies for domestic animals in *The Variation of Animals and Plants under Domestication.* Specific evidence for the same relationship in the human female comes from two sets of published data: (1) Chronically undernourished girls had menarche at the same mean weight as well-nourished girls of the same ethnic stock, but two years later than the well-nourished girls, and (2) the mean weight at menarche (about 47 kilograms) has not changed in western Europe or the U.S. in the past 125 years, but the mean age of attaining that weight has fallen from 16.5 to 12.8 years.

These observations on the relation of weight and sexual maturation do not exclude the possible modulating influences of other factors for man or animals, such as stress, temperature, pheromones or, as was suggested in 1790, sweet music. Dr. M. A. Raciborski, *lauréat* of the Academy of Medicine in Paris, reported in 1844 an admirably direct experiment in which members of the Paris orchestra were brought to the Paris zoo and placed behind a curtain at the elephant cage in order to test whether music hastened sexual maturation. The orchestra played selections from Gluck's *Iphigénie en Tauride,* while the reactions of the two elephants were carefully noted. The experimental results were positive: the elephants reacted passionately, and they matured eight years earlier than elephants in the wild. An alternative explanation is that elephants, like other mammals, have a weight-dependent sexual maturation, and elephants in zoos are better fed and move around less than those in the wild.

ROSE E. FRISCH, PH.D.

Center for Population Studies
Harvard University
Cambridge, Mass.

ERRATUM

In "The Cell Cycle," by Daniel Mazia (SCIENTIFIC AMERICAN, January), it is stated: "Max M. Burger of Princeton University was the first to find that the membrane of contact-inhibitable cells reacts with those plant proteins that induce cycling in noncycling lymphocytes.... Cells that are not contact-inhibitable are unaffected by the presence of the protein."

The sentences should read: "Max M. Burger of Princeton University was the first to find that the membrane *of cells that are not contact-inhibitable* reacts with those plant proteins that induce cycling in noncycling lymphocytes.... Cells that *are* contact-inhibitable are unaffected by the presence of the protein."

Scientific American, May, 1974; Vol. 230, No. 5. 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, $12 ☐ 2 years, $22 ☐ 3 years, $30
All other countries:
☐ 1 year, $15 ☐ 2 years, $27 ☐ 3 years, $37

NAME

NEW ADDRESS

OLD ADDRESS

"My new BMW gives me a better kind of luxury. High performance with excellent gas mileage. That's great engineering."

These remarks from Senior Airlines Captain Stanley Adelman, of Westport, Connecticut, are typical of the feelings many BMW owners have for their cars.

"As a veteran commercial pilot, I have profound respect for the truth that form follows function. For example, a jet aircraft looks sleek and clean-lined because these factors contribute to its performance.

"And exactly the same principles apply to my 3-liter BMW. The car is fast, maneuverable, and amazingly sure-footed. Its controls are light and very positive. And its aerodynamics were proven in wind-tunnel tests.

"And incidentally, I regularly get over 20 miles to a gallon.

"I've owned a number of so-called luxury cars. But I think my BMW is the best-engineered car in the world."

At BMW, we build cars that above all are more *efficient.*

Road & Track calls BMW's new 3-liter engine the most sophisticated in-line six in the world. Smooth, dependable, and brilliantly responsive, yet remarkably easy on fuel.

Our fully independent four-wheel suspension gives BMW superior roadholding capabilities.

Extraordinary handling characteristics are but one of BMW's own demanding safety criteria. BMW safety features are built into the car's design, not just tacked on.

The fast, rugged, spirited 3-liter BMW is designed to achieve technological superiority in every aspect of automotive construction. In the highest tradition of the Bavarian Motor Works, it is more than a fine car. It is what luxury *should* be—the luxury of great engineering.

BMW automobiles are available in a full range of two and four door models.
Sole U.S. Importer: **Hoffman Motors Corporation.**
375 Park Avenue, New York, N.Y. 10022;
12541 Beatrice Street, Los Angeles, Calif. 90066.
Dealers Coast to Coast.
Advantageous European Delivery Program.
Write for free brochure. Code No. 01

Canada: East: BMW Distributors Eastern Canada, Ltd. Ontario.
West BMW Distributors (Western) Co.
British Columbia

Bavarian Motor Works

BMW

20 ways to on your

1 Avoid paying extra installation charges when you order new phone service. Have all the work done at one time. Changing your mind later will mean extra visits and extra charges. So consider carefully all the different colors and styles, how many phones you want and exactly where you want them installed.

2 Ask one of our business office Service Representatives to explain the different types of service offered by your local Bell Company. Choose the one which best fits your pattern of calling. If you don't do a lot of calling each month, ask if "budget" or "limited" service is available in your area at a lower rate.

3 Ask the Service Representative for a rundown on the specific rates and charges you can expect. Find out exactly what the regular monthly charges will be, and what the one-time-only payments are. Find out whether there are options in your area of paying on a monthly basis, or making a single one-time only payment, on certain items.

4 Are you going to be away from home for any extended period of time? A business office Service Representative can tell you, based on how long you plan to be away, whether you could save money by temporarily suspending your telephone service.

5 Moving to a new residence? Ask a Service Representative whether you are eligible for a credit on your bill if you take your present phones along with you to your new location.

6 If you've never had a phone in your name before, or have never established credit, you may be asked to pay a deposit when you order telephone service. But we don't like to keep deposits for long periods of time. In fact, if you establish good credit with us by paying on time, we'll return your deposit to you PLUS interest. Ask a Service Representative to explain the details which apply in your area.

7 If a coin phone swallows your money but doesn't give you your call, you're entitled to a refund. Find a phone that works, dial "Operator" and explain what happened. You'll get a refund in the mail. P.S. We'd appreciate it if you'd also tell the operator the telephone number and location of the phone that's out of order so we can get it fixed as soon as possible.

8 Reach a wrong number on a Long Distance call you just dialed? Don't just hang up. Ask for the area code and the number you reached in error. Then dial "Operator" and report what happened. The operator will have the charge removed.

9 Get a poor connection on a Long Distance call, or get cut off in the middle of your conversation? Don't just hang up and call back. The person who placed the call should report what happened to an operator. The operator will issue a credit for the time your call was interrupted.

10 Error on your bili, with a charge for a Long Distance call you didn't make? Call the business office. A Service Representative will arrange to get the charge removed.

11 Save on Long Distance charges by cutting down on person-to-person calls. It's true you may not be able to reach the person you want on your first try with a station-to-station call. But in many instances you can make two (or even three) out-of-state station-to-station calls for what it would cost you to make that one person-to-person call. This is particularly true if you dial your own calls instead of going through an operator.

12 Dialing your own out-of-state Long Distance calls is the least expensive way of all. If you don't know the number for a call you want to make to a distant city, you can obtain it at no charge to you by dialing the area code (when required) for that city, plus 555-1212, for Directory Assistance. Then dial direct and save. Save time

save money phone bill.

in the future by recording the number in your personal number book. A listing of all area codes can be found in the information pages at the front of your local telephone directory.

13 Make sure you know when dial-direct rates apply before you make your call. They apply on all out-of-state calls to anywhere in the United States (excluding Alaska) if they are completed from a residence or business phone without an operator's assistance. They also apply on calls placed with an operator from a residence or business phone when direct dialing facilities are not yet available.

14 But it's even more important to know the circumstances when direct-dial rates do NOT apply. They do not apply on person-to-person, hotel-guest, credit card or collect calls, or on calls charged to another number, because an operator must assist on such calls. Direct-dial rates do not apply on calls made from coin phones, even those from which you dial the complete number yourself before the operator comes on.

15 While operator-handled calls cost you more than those you dial yourself, there is one exception. If you run into equipment trouble completing a Long Distance call you're dialing yourself from a home or business phone, you're still eligible for the dial-direct rate even if you require an operator's assistance. Explain your problem to an operator. If you need help in getting the call through, or in making a satisfactory connection, confirm with the operator that it will be charged at the dial-direct rate.

16 Dial direct, but save even more by making your out-of-state Long Distance calls within the time periods when rates are lowest. The lower rates for out-of-state calls made in the evening, on the weekends or late at night are described for you in the call guide in the front of your local directory.

17 Don't get caught by surprise, or miss out on the money you can save on Long Distance calls you make within your state. The times when lower rates apply may be different than for your out-of-state calls. Check carefully in the call guide in the front of your local directory for a description of when to save on calls you make within your state.

18 If you're concerned about avoiding added charges on your Long Distance calls in general, don't guess how long you've been talking. Time yourself, so you can finish your call before overtime rates apply. To save even more time and money, jot down what you want to say before you dial.

19 Before you go ahead and place a Long Distance call to a business, check first to see if they have a toll-free number. You can recognize it because it has an 800 prefix instead of a regular area code. If they have one, it's usually displayed in their advertising, or you may find it listed in your own local telephone directory. If so, the call's on them, and you save.

20 The information pages at the front of your local telephone book are a good source for tips on how to place your calls and how to save time and money in using your telephone. Look in the book, and save.

We hear you.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1924: "The atomic weight of hydrogen is not exactly 1 but by careful measurement is found to be 1.0077. In this slight discrepancy an immense store of possible energy is indicated, which someday, when we have learned how, may become accessible for good or ill to the human race. If the whole of any perceptible portion of matter disappeared, the energy resulting would be prodigious, being multiplied by a factor equal to the velocity of light. When hydrogen is packed into helium, the whole runs not the slightest risk of disappearing, but seven or eight parts in every 10,000 do disappear. The 1.0077 becomes 1.0. And though the disappearing fraction is small, the total of which it is a fraction is so gigantic that the result would put all our other sources of energy to shame. If ever the human race should get hold of a means of tapping the energy contained in the atom, the consequences will be beneficent or destructive according to the state of civilization at that time attained."

"The basis of nervous activity is formed by the so-called reflexes or instincts. Food, for example, stimulates the food reflex, which consists of certain movements in animals and the secretion of saliva. Professor I. P. Pavlov of Petrograd reports that if some indifferent agent, which previously had nothing in common with feeding, is repeated many times with the feeding, after a time it begins to stimulate the food reflex when used alone. Thus if we produce some distinct musical sound at a given frequency and always at the same time feed a dog, after a while this sound, used alone, will produce the same food reaction as the food itself. In this way, besides the reflexes or instincts that are inborn, some reflexes are acquired during the life of the individual. The inborn reflexes we call unconditioned and the acquired reflexes we call conditioned."

"In 1899 Sir William Crookes pointed out that the world's food supplies are dependent on a supply of nitrogenous fertilizers to the soil. Each crop takes so much out of the soil that unless this essential material is replaced the yield per acre steadily drops. There is an inexhaustible store of nitrogen in the air, but nitrogen as such is one of the most inert of materials. It is only when it has been made to combine with other elements, such as hydrogen, that it becomes available as plant food. The problem consists in taking from the two abundant sources, air and water, the constituent nitrogen and hydrogen and combining them to form a new substance, ammonia. At the beginning of the present century there was no known method for any such fixation of nitrogen, but as a result of applying pure research there are now many industrial methods. The first problem was to find the conditions under which the equilibrium point of the reaction would be most suitable. This was done when it was found that the production of ammonia is increased by the application of 200 and even 1,000 atmospheres of pressure. The ammonia must be produced at a practical rate and to bring this about substances known as 'catalyzers' are employed. By applying these principles the Germans have succeeded in producing ammonia at the rate of more than 1,000 tons a day."

SCIENTIFIC AMERICAN

MAY, 1874: "Babbage, in speaking of his analytical engine, has suggested that a machine might be made that would play a game of combination, such as draughts, provided the maker of the machine himself would work out perfectly the sequences of the game. Professor Fairman Rogers finds that the sequences of tit-tat-toe are easily tabulated, and hence an automaton may be made that will play the game as follows: The opponent to the automaton makes the first move in the game and in so doing causes a certain cylinder to change its position. This causes the automaton to make that play which the proper sequence of the game requires. The next play of the opponent moves another cylinder, and so on throughout the sequence. If the player plays perfectly, the game will be drawn; if the opponent makes a mistake, the automaton will take advantage of it and win the game."

"The bicycle, after going entirely out of fashion as a toy, is now being put to some practical use. Messengers, called 'véloce men,' thus mounted convey dispatches in Paris from the Bourse—or stock exchange—to the central telegraph bureau. The distance is about six miles and is accomplished in 25 minutes, at a charge of 50 cents. A company is being formed to place a very large number of velocipedes on the streets and to supply messengers to go to any part of the city. The Parisian journals are also using the bicycle to obtain quick reports. The distance from the palace of Versailles to Paris—about 13 miles—can be made in 45 minutes, faster than the ordinary trains on the railroad."

"Experiments performed by Professor Arthur W. Wright of Yale College will probably set at rest the moot question as to the nature of the zodiacal light. This faint nebulous radiance is seen in certain seasons, especially in the Tropics, in the west after twilight has ended or in the east before it has begun. Professor Wright, employing a new apparatus of his own design, has determined that the light is polarized in a plane passing through the sun and thus has proved it is reflected and derived originally from the sun. Confirming this view, it was found that the spectrum of the light is the same as that of the sun, except in intensity. Professor Wright concludes that the light is reflected from matter in the solid state, from innumerable small bodies revolving around the sun in orbits, of which more lie in the neighborhood of the ecliptic than near any other plane passing through the sun."

"The question of the disposition of the dead by burning the bodies, after the manner of the ancients, has for some years been under discussion in scientific circles in Europe. The reader unversed in the process will naturally think of the ancient pyre and will probably suppose that it is the intention of the advocates of the plan to burn the bodies on huge piles of variously scented wood. Little would be gained in an aesthetic or even a sanitary point of view if such were the system, for the gases and fumes evolved would be far from healthful. The body of the poet Shelley was thus destroyed, and his biographer tells us that, so far from being the beautiful and poetic rite intended, the process was a very disagreeable and nauseous operation. Science provides a better plan for reducing ashes to ashes: an apparatus devised by Professor Brunetti especially for the purpose. It consists of an oblong furnace lined with fireproof brick. By means of this apparatus, at the expense of about 150 pounds of wood, complete cremation can be effected in two hours."

Energy.

Good help is hard to find, so the saying goes, and that's a stark reality when it comes to our greatest helper, energy.

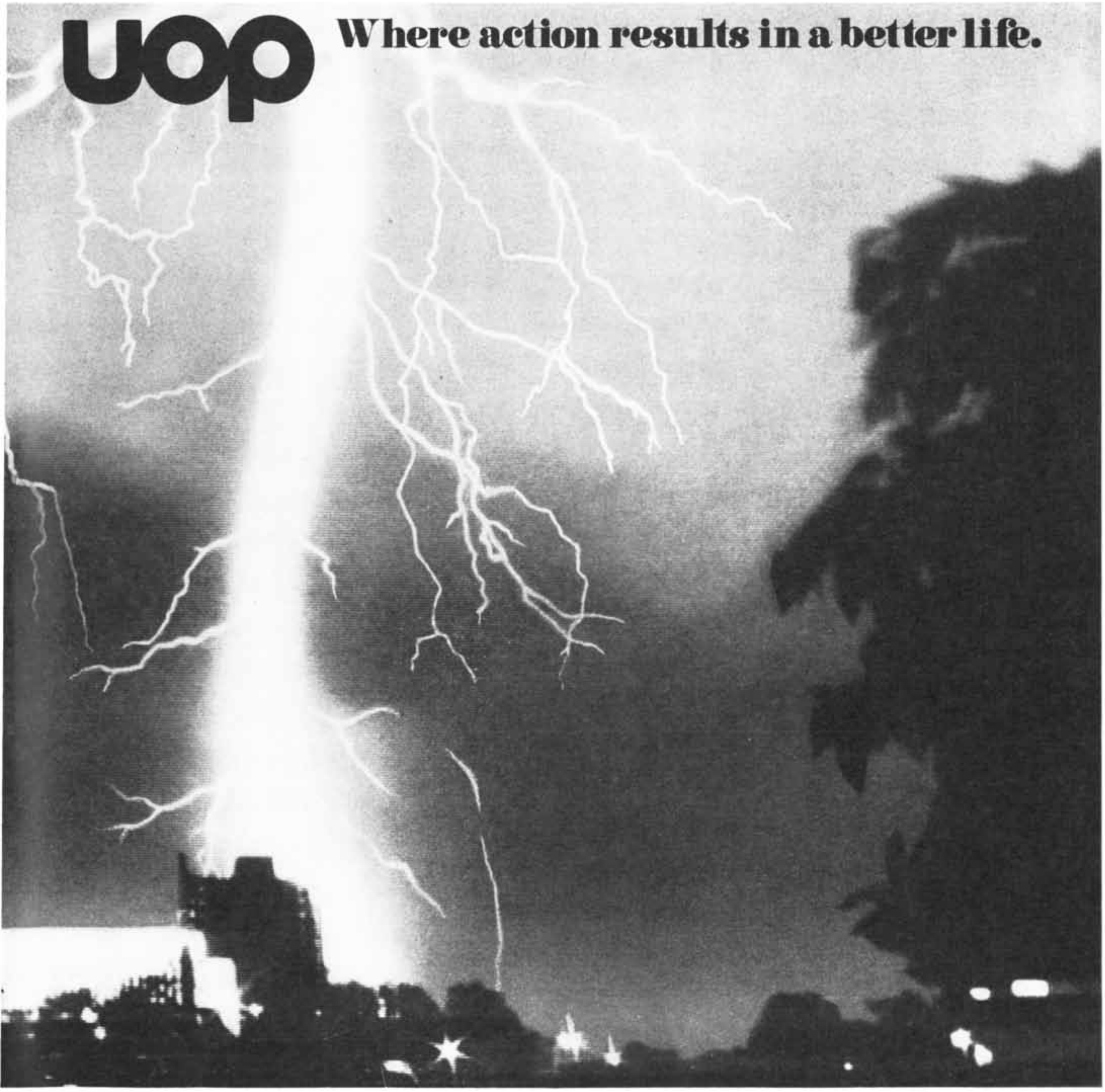
We're in the business of helping people make better use of energy, wherever and whenever it's found. And we have been, ever since petroleum was first "cracked" from crude, some sixty years ago. (One of our founders patented that process, and we now own some 10,000 unexpired patents, most dealing in petroleum science and petrochemicals.)

UOP processes are designed for obtaining more usable energy, more efficiently, from available resources. Including grass roots

refineries. With or without an energy problem, there's still going to be a widespread need for that expertise.

We're ready to meet that need, with a deep, broad team of technology management. And with a background of sixty years' experience in research and development, including the fathering of Platforming® and 34 other licensed processes.

Energy is an entire industry for UOP—and also a matter of a better life. Our energy projects are investments in tomorrow. To obtain more information about our company and sharing in that investment, write UOP, Des Plaines, Illinois 60016.



THE AUTHORS

BARRY E. CARTER ("Nuclear Strategy and Nuclear Weapons") practices law in Washington as an associate of the firm of Wilmer, Cutler & Pickering. Before joining the firm in 1972 he had worked for several years as a Government official concerned with international affairs. For a year starting in 1969 he was a program analyst in the NATO and General Purpose Division of the Office of the Assistant Secretary of Defense for systems analysis. From October, 1970, to June, 1972, he was on the staff of the National Security Council, working mostly on the Strategic Arms Limitation Talks. From June to December of 1972 he was a research fellow in the Institute of Politics of the John Fitzgerald Kennedy School of Government at Harvard University. Carter was graduated from Stanford University in 1964. He obtained his master's degree in economics and public affairs at the Woodrow Wilson School of Public and International Affairs of Princeton University in 1966 and his law degree from the Yale Law School in 1969.

MARTIN BIDDLE ("The Archaeology of Winchester") is director of the Winchester Research Unit of the Winchester Excavations Committee. "My main interests," he writes, "lie in two areas: the formation of medieval Europe from the aftermath of the Roman empire in the West, that is, the period from the third century to the 12th century, and the introduction of Renaissance architectural and artistic ideas to northern Europe and particularly England in the first half of the 16th century." He is also involved in efforts to prevent the destruction of archaeological evidence by "trying to get governments to take action before events have removed the future from our past." Biddle received his bachelor's and master's degrees from the University of Cambridge in 1961 and 1965 respectively and his second master's degree from the University of Oxford in 1967. From 1961 to 1963 he was Assistant Inspector of Ancient Monuments in the Ministry of Public Building and Works. He then spent four years as lecturer in medieval archaeology at the University of Exeter and one year as a visiting fellow at All Souls College, Oxford.

R. W. GUILLERY ("Visual Pathways in Albinos") is professor in the anatomy department of the University of Wisconsin Medical School. Born in Germany, he was educated in England, receiving his bachelor's degree and Ph.D. at University College London in 1951 and 1954 respectively. He taught anatomy at University College London for 11 years before moving to the University of Wisconsin in 1964. Guillery wishes to acknowledge the assistance of Grayson Scott, Elaine Langer and Barbara Yelk and grants from the U.S. Public Health Service.

FERREN MACINTYRE ("The Top Millimeter of the Ocean") is a Dyason Fellow in Environmental Studies at the University of Melbourne. His degrees are in chemistry, a subject to which he turned after a certain amount of exploring. "I worked for eight years in the real world," he writes, "as lumberman, machinist, machine designer and loftsman before retiring to college." He obtained his bachelor's degree at the University of California at Riverside in 1960 and his Ph.D. from the Massachusetts Institute of Technology in 1965. He has worked at the Scripps Institution of Oceanography and the Marine Science Institute of the University of California at Santa Barbara.

NATHAN SHARON ("Glycoproteins") is a professor at the Weizmann Institute of Science in Israel. He recently became head of the department of biophysics in succession to Ephraim Katzir-Katchalski, who is now president of the State of Israel. Sharon's avocations include writing popular science and swimming. He writes that he is "happily married to a schoolteacher, and both of our grown-up daughters would like to choose science as their career."

DOUGLAS SCOTT, WILLIAM W. SEIFERT and VERNON C. WESTCOTT ("The Particles of Wear") are respectively head of the metallurgy section of the National Engineering Laboratory in Scotland, professor of electrical engineering and of engineering and civil engineering at the Massachusetts Institute of Technology and chairman of the board of Trans-Sonics, Inc. Scott, who is editor of the journal *Wear*, was with the Ministry of Supply for 10 years before joining his present laboratory in 1950. Seifert, who was graduated from Rensselaer Polytechnic Institute in 1941, has been at M.I.T. since going there as a graduate student. He received his master's degree in 1947 and his Sc.D. in 1951. Westcott was graduated from R.P.I. and was employed for several years by the Raytheon Manufacturing Company until in 1948 he founded (with Edward T. Rigney) Trans-Sonics, which he served as president from 1948 to 1960. The authors wish to acknowledge support by the Office of Naval Research and the Advanced Research Projects Agency of the U.S. Department of Defense.

DONALD P. GREENBERG ("Computer Graphics in Architecture") is professor in the department of architecture at Cornell University and director of the university's program of computer graphics. He was graduated from Cornell in 1958 and took his Ph.D. there (in structural engineering) in 1968. Greenberg wishes to thank the General Electric Corporation for enabling him to use its Visual Simulation Laboratory and to acknowledge assistance from R. Rougelot, E. E. Carey, W. Hartmann and W. Gannon of General Electric and David Simons, Robert Hastings, William Cunningham, Thomas Fridstein, Marc Levoy, Nicholas Lindabury, David McNeil, David Montanari, John Nicolls, Alfreda Radzicki, David Ross, Stephen Snyder, Diego Suarez-Betancourt and Nicholas Weingarten of Cornell.

JAY M. PASACHOFF and WILLIAM A. FOWLER ("Deuterium in the Universe") are respectively at Williams College and the California Institute of Technology; Pasachoff is director of the Hopkins Observatory and assistant professor of astronomy and Fowler is Institute Professor of Physics. Pasachoff was graduated from Harvard College in 1963 and also received his master's and doctor's degrees at Harvard. Fowler obtained his bachelor's degree in engineering physics at Ohio State University in 1933 and then went to Cal Tech as a graduate student. Except for periods as a Guggenheim Fellow at the University of Cambridge and as a visiting lecturer at a number of institutions in the U.S. and Europe, he has remained at Cal Tech ever since. He received his Ph.D. there in 1936, spent three years as a research fellow and then was appointed to the faculty. When he won the $25,000 Vetlesen Prize in 1973 "for outstanding achievement in the sciences resulting in a clearer understanding of the earth, its history, or its relation to the universe," the citation read: "It is probably fair to say that almost all of our quantitative information about the basic nuclear processes that enter into stellar energy generation and element synthesis is due to Fowler or to work directly instigated by him."

Teach your students on the equipment they'll probably use as professionals.

When it comes to providing practical job-training, many technical educators choose Hewlett-Packard test and measuring equipment for their classrooms and training labs. Because of its excellent performance at fair prices HP equipment is widely used in most technical disciplines—practically every employment situation your students are likely to encounter.

Of the more than 3,000 products manufactured by HP, a number are well-suited for classroom use. These include power supplies, oscilloscopes, voltmeters, counters, signal sources, recorders, and calculators. Complete descriptions of these instruments are brought together in our new catalog, "Electronic Test and Measuring for Technical Education." This catalog is free to educators for the asking.

A Hands-On Way to Learn Digital Electronics

One of the educational instruments featured in our catalog is the new HP Logic Lab. Designed for hands-on instruction, the Logic Lab is a complete course in digital electronics including textbook, lab workbook and a fully equipped portable laboratory.

Your students learn to assemble a variety of functioning digital circuits using the breadboard, components and interconnecting wiring supplied. With the Logic Lab, circuit assembly is no longer a grueling process of soldering and un-soldering IC's and wires because the student uses solderless connection techniques for assembling components and wires on the plug-in breadboard.

To make his circuit come to life, the student simply places the breadboard in the laboratory station mainframe. All the essentials needed to stimulate the circuits are built-in: power supply, logic state indicators and programmers, and pulse sources.

Now the student can verify his circuit's operation using professional quality digital test instruments—the HP Logic Probe, Logic Pulser and Logic Clip. In doing so, he not only learns how to troubleshoot a digital circuit, but also sees visual evidence of the behavior of the components and can make sense of their interactions. In the final analysis, this is where he learns what the 1's and 0's mean.

Even if your student is caught by the bell in the middle of assembling or testing a circuit, the Logic Lab is always ready for the next class. One student's breadboard is easily removed and replaced by another without disturbing the circuits.

Written by a team of experienced educators and industry engineers, the Logic Lab text is arranged in a modular format that results in an effective self-learning package. Your student can start on a level and learn at a pace that suits his individual needs.

The complete Logic Lab package is available for $650.*

The first fully programmable pocket-sized computer calculator.

The new HP-65 does what many still find hard to believe of an 11-ounce instrument that fits in the palm of your hand. It not only makes *single* complex calculations but also performs the complete *sequence* of calculations necessary to solve a problem. All you do is enter the raw data and a pre-recorded program. The HP-65 remembers and executes the correct sequence of keystrokes for solving the problem and gives you the answer automatically.

The program is recorded on a magnetic card the size of a small stick of gum. Each program can consist of as many as 100 steps, and each step can be any of the calculator's 51 arithmetic, logarithmic, trigonometric and other transcendental keystroke functions. These can be combined with branching, logical comparison and conditional skip instructions, using the HP-65's four-register operational stack and nine addressable memory registers.

You can write, edit and record your own programs on the HP-65. It couldn't be easier; push the keys in the same sequence as you would use to solve the problem manually, with the programming switch in the W/PRGM position. Then switch to WRITE to record these steps on the magnetic card. Next time

you want to solve the same problem, turn the switch to RUN and insert the card.

If you don't want to write your own programs, use ours. Six pre-recorded packages, called Application Pacs, are available now; more are on the way. Each Pac contains up to 40 programs in disciplines such as mathematics, statistics, surveying, medicine, and electrical engineering. The Math Pac I, for example, includes pre-recorded programs for hyperbolic functions, first order differential equations, solving simultaneous equations in three unknowns and 37 other calculations. Each Pac costs $45.*

The HP-65 is priced at $795* including one standard Application Pac with representative programs and blank cards, handbook, battery pack, AC adapter/recharger, carrying case and travel case.

For more information on these products write to us. Hewlett-Packard, 1504 Page Mill Road, Palo Alto, California 94304.

*Domestic USA prices only. 00411

Mail to: Hewlett-Packard, 1504 Page Mill Road, Palo Alto, CA 94304

Please send me information on: ☐ HP 65 Computer/Calculator
☐ "Electronic Test and Measuring for Technical Education" Catalog
☐ HP Logic Lab

Name________ Title________

Company________

Address________

City________ State________ Zip________

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.

Palo Alto, California 94304

Established 1845 SCIENTIFIC AMERICAN May 1974 Volume 230 Number 5

Nuclear Strategy and Nuclear Weapons

The "improved counterforce capability" proposed by the Administration is viewed as not only unnecessary and potentially costly but also likely to undercut the SALT negotiations and increase the risk of nuclear war

by Barry Carter

"Should a President, in the event of a nuclear attack, be left with the single option of ordering the mass destruction of enemy civilians, in the face of the certainty that it would be followed by the mass slaughter of Americans? Should the concept of assured destruction be narrowly defined and should it be the only measure of our ability to deter the variety of threats we may face?"

The questions asked in the preceding quotation, taken from President Nixon's first foreign-policy report in 1970, have been cited repeatedly in the past few months by Administration spokesmen in an effort to explain and justify some significant changes that are being made in U.S. policy regarding its strategic military forces. The new strategy, spelled out most clearly in Secretary of Defense James R. Schlesinger's annual report for the fiscal year 1975, released in March, seeks "to provide the President with a wider set of much more selective targeting options," and hence greater "flexibility," in choosing an appropriate response to "any kind of nuclear attack."

As the opening quotation illustrates, much of the official rhetoric concerning this new development in U.S. strategic policy has been more misleading than illuminating. To criticize the "assured destruction" doctrine of the past decade or so as planning only for massive retaliation against Russian cities ignores the fact (belatedly acknowledged by Schlesinger) that U.S. strategic forces have for years had the capability, both in weapons and in planning, for a "flexible response." More important, the broad hypothetical issues invoked by such public statements have tended to obscure the more immediate real issues presented by this Administration's recent actions.

The real issues are serious ones. The primary operational question at present is whether or not the U.S. should develop missiles with an improved capability for attacking "hardened" targets in the U.S.S.R. The main rationale offered for developing such an improved "counterforce" capability (so called because it is aimed at an opponent's military forces) is that it is "impermissible" for the U.S. not to "match" certain Russian counterforce developments. There is also the suggestion that these missiles would minimize "unintended collateral damage."

The preceding question in turn raises the subtler issue of how the active promotion of such programs for improved counterforce capabilities affects the stability of the strategic nuclear deterrent and hence the likelihood that there will be a nuclear war. Before one can address these two issues one must understand why public debate should properly focus on such questions and not (at this time anyway) on the kind of questions posed in President Nixon's 1970 remarks.

In the late 1950's and early 1960's U.S. strategic policy went through a series of transformations. By 1962 American military planners recognized that the U.S. would have many more missiles than the U.S.S.R. could have for several years and in fact many more missiles than were required to devastate every major city in the U.S.S.R. A counterforce strategy therefore held out the attractive option of limiting damage to U.S. cities by destroying a substantial part of the Russian strategic forces. In language that sounds remarkably familiar today, Secretary of Defense Robert S. McNamara said in a speech in Ann Arbor, Mich.: "The United States has come to the conclusion that, to the extent feasible, basic military strategy in a possible general nuclear war should be approached in much the same way that more conventional military operations have been regarded in the past. That is to say, principal military objectives, in the event of a nuclear war stemming from a major attack on the alliance, should be the destruction of the enemy's military forces, not of his civilian population."

The Russians, however, continued to deploy land-based intercontinental ballistic missiles (ICBM's) and submarine-launched ballistic missiles (SLBM's). As a result, even if the U.S. sought to limit damage to itself by the partial destruction of the Russian strategic forces, there would still be more than enough Russian forces left to kill tens of millions of Americans. Recognizing this fact, McNamara increasingly emphasized by the mid-1960's the concept of "assured destruction," which he said in 1968 meant the "ability, even after absorbing a well-coordinated surprise first strike, to inflict unacceptable damage on the attacker." This criterion he defined explicitly: "In the case of the Soviet Union, I would judge that a capability on our part to destroy, say, one-fifth to one-fourth of her population and one-half of her industrial capacity would serve as an effective deterrent."

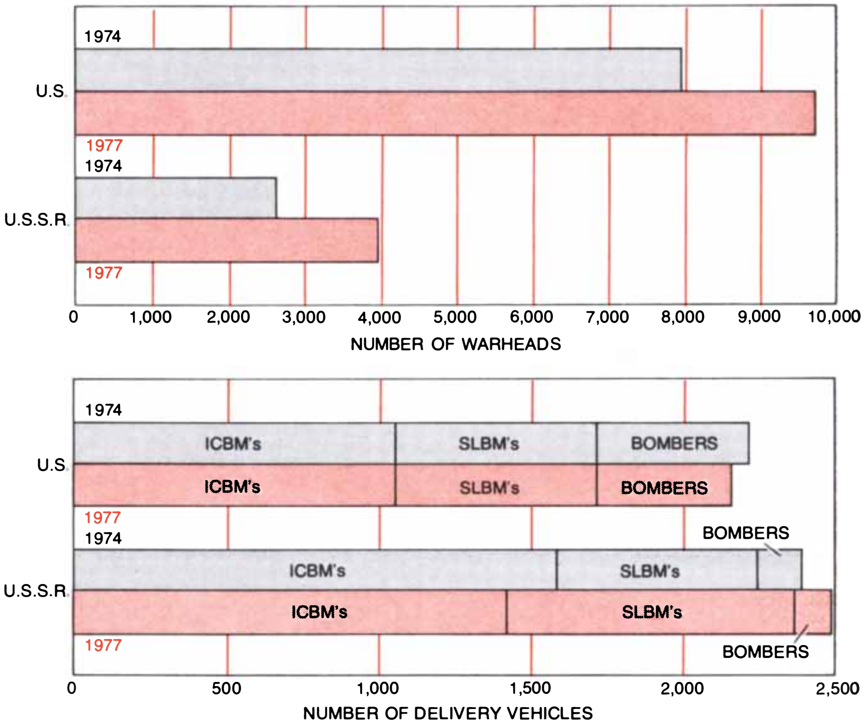
Few concepts have been as maligned or misunderstood as that of assured destruction. Critics label it genocide or use the acronym of "mutual assured destruction" to call it MAD. In fact, the concept seems well designed to serve two purposes. First, by planning the size of U.S. forces on the basis of the "worst case" scenario of an all-out Russian surprise attack, it ensures that the U.S. possesses the ultimate threat: to be able to wipe out the U.S.S.R. or any attacker in retaliation. Second, since the destruction criterion is reasonably precise, the concept provides a useful basis for limiting strategic-weapons procurement and for evaluating arms-control proposals.

While retaining the assured-destruction concept, McNamara and his successor, Clark Clifford, supervised the development of the wide array of weapons that constitutes today's U.S. strategic arsenal. Both the numbers and the characteristics of many of these weapons were consistent with the assured-destruction concept, partly because the U.S. possesses a "triad" of strategic offensive forces and partly because of the hedge against the "highest expected threat." The triad approach seeks to maintain a major retaliatory capability in each component of our strategic offensive forces: ICBM's, SLBM's and long-range bombers. Justified on the grounds that each component presents a different problem for an attacker, difficult and costly problems for his defense and a hedge against unexpected failures in one or both of the other components, the net result of the triad approach is to provide in the aggregate a high degree of confidence that the assured-destruction mission could be carried out.

The hedge against the highest expected threat, as projected in the National Intelligence Estimates, meant that weapons would be developed and sometimes procured as a cushion against Russian developments that, although not considered likely, were possible. The predictable result was that the U.S. came to possess much more powerful forces than were shown by subsequent events to be required for assured destruction. For example, one of the main justifications offered for developing multiple independently targeted reentry vehicles (MIRV's) was to hedge against a greater-than-expected Russian deployment of an anti-ballistic-missile (ABM) system, on the theory that increasing the number of incoming warheads would enable the U.S. offense to penetrate the Russian defense more easily.

Of course, some of the development and procurement decisions also reflected inevitable political and bureaucratic pressures. For example, faced with pressures from the military and from Congress, McNamara apparently thought he could not ask for fewer than 1,000 Minuteman ICBM's.

Finally, the proponents of the assured-destruction concept in the latter half of the 1960's quietly subscribed to secondary strategic objectives, in particular the desire to retain some ability to respond flexibly in the case of an actual attack. If the U.S. were subjected to a "limited" nuclear attack—possibly with a small number of missiles or because of an accidental launch—most thought the President should have a range of options from which to choose. This factor helps to explain why, for example, the Minuteman II warhead, which was first deployed in 1966, could be programmed for up to eight alternative targets, and why there was flexibility in the actual targeting plans.



STRATEGIC NUCLEAR ARSENALS of the U.S. and the U.S.S.R. are compared here in terms of two key indicators: the number of individually targetable nuclear warheads, including bombs (*top*) and the number of delivery vehicles (*bottom*). Each pair of bars represents the total operational strategic forces projected for both sides as of mid-1974 (*gray*) and 1977 (*color*), based on data provided by the U.S. Government. The delivery-vehicle bars are broken down into segments representing land-based intercontinental ballistic missiles (ICBM's), submarine-launched ballistic missiles (SLBM's) and long-range bombers. As the bars show, the U.S. is at present either ahead of or in rough parity with the U.S.S.R. in these two categories and is expected to remain so for at least the next few years. This purely numerical representation of the military balance between the two superpowers does not, of course, take into account such important factors as qualitative differences in weapons technology, the larger size of the Russian ICBM's or the thousands of U.S. "tactical" nuclear weapons in Europe, Asia and aboard forward-based ships. More important, regardless of the exact numerical relation of the two arsenals, each side now has, and for the foreseeable future will continue to have, a secure retaliatory capability, that is, more than enough strategic forces to absorb even an all-out nuclear attack by the other side and still be able to retaliate by carrying out a wide variety of limited nuclear attacks or by inflicting an unacceptable level of "assured destruction" on the population and industry of the attacker.

As a result the U.S. ended up with strategic-war capabilities considerably greater than the assured-destruction concept required. That this situation was rarely acknowledged publicly was a serious mistake, the results of which we are now reaping in public misunderstanding of the policies of the past and, more important, in the sometimes surprising ignorance about the present capabilities of the U.S. strategic forces. The simple fact, which cannot be stressed too strongly, is that the U.S. strategic forces are now capable of carrying out a large array of alternative missions, far in excess of assured destruction.

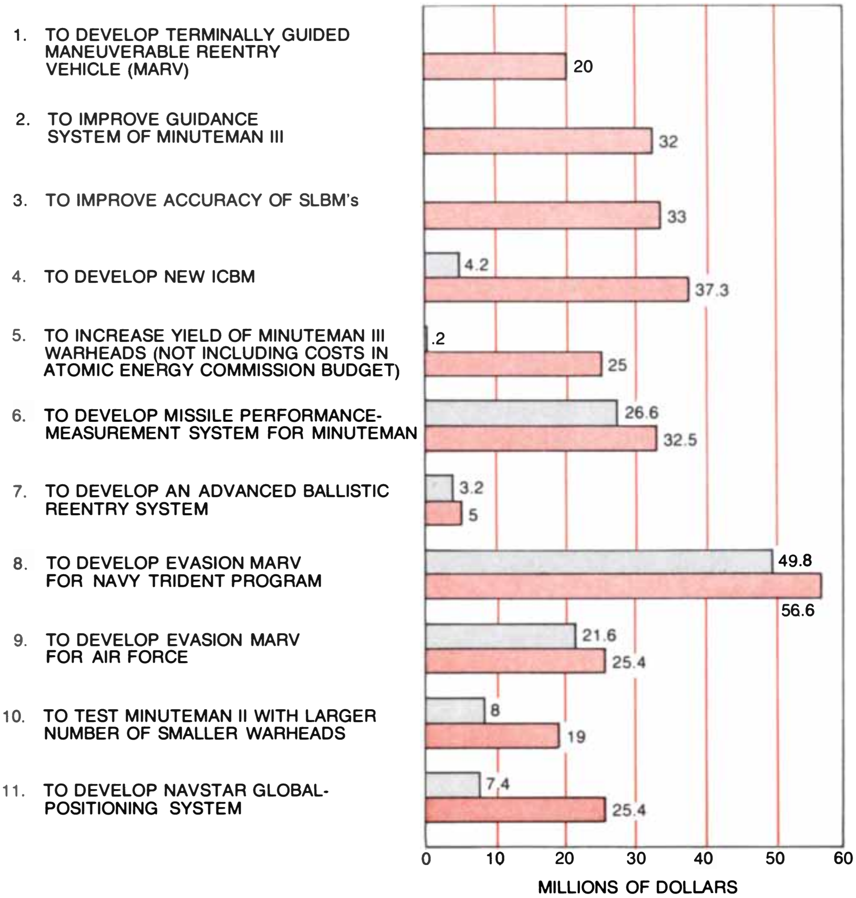
To begin with, assured destruction does not require many forces. Assuming zero or low Russian ABM levels (a reasonable assumption given the 1972 Moscow Treaty limiting ABM systems), the delivered warheads of 220 Minuteman III ICBM's could kill about 21 percent of the Russian population from immediate effects alone and destroy about 72 percent of the Russian industrial capacity. The delivered warheads from 170 Poseidon missiles (which is fewer than the total carried by 12 submarines) could cause a similar level of damage [*see illustration on page 24*]. Projections of bomber survivability vary greatly, but most experts would estimate that enough B-52's could reach their targets to satisfy easily the traditional assured-destruction criterion.

The total of U.S. strategic forces is, of course, much larger. There are at present 1,054 ICBM's, of which 1,000 are Minuteman missiles and 54 are the older, larger Titans. Of the Minuteman missiles 550 have been or are in the process of being converted to the Minuteman III, which can carry up to three warheads. These MIRV's are estimated to have an accuracy of 1,500 feet or less (expressed in terms of "circular error probable," which means that 50 percent of the warheads are expected to fall within a radius of 1,500 feet of the target). The explosive power, or yield, of each warhead is equivalent to between 170 and 200 kilotons of TNT, or at least 11 times the size of the 15-kiloton bomb dropped on Hiroshima. Rapid retargeting of the Minuteman III will be possible soon with the advent of new computer-software systems, such as the Command Data Buffer system. (All estimates of the numbers and characteristics of U.S. forces used in this article are taken from the statements of U.S. officials, from publications of the International Institute of Strategic Studies and from other reliable publications.)

In addition the U.S. arsenal includes 656 SLBM's, 496 of which are scheduled to become Poseidon missiles. The Poseidon can carry up to 14 MIRV's, but it is usually deployed with 10. Although accuracy might be reduced by uncertainties about the submarine's location, it still is probably less than 3,000 feet. Moreover, even though each warhead is smaller than Minuteman's, there are many more of them and each is still about three times the size of the Hiroshima bomb. Like the Minuteman III warheads, the Poseidon warheads can be retargeted quickly.

Bombers are often viewed as the stepchild of the U.S. strategic triad. The approximately 400 B-52's and 65 FB 111's are unaccountably ignored in many comparative tables of American and Russian strategic forces, notably in President Nixon's first three foreign-policy reports. This is surprising given the fact that an estimated 40 percent of the U.S. budget for strategic offensive forces is spent on bombers. Moreover, from the standpoint of nuclear strikes the per-sortie attrition rate of about 3 percent suffered by the B-52's in their attacks on heavily defended Hanoi demonstrated high survivability. Indeed, most places in the U.S.S.R. would not be as heavily defended as Hanoi, the B-52's would not be making the more vulnerable high-altitude attacks they made there and the bombers would use nuclear warheads to silence air-defense batteries. Each B-52 carries between four and 24 nuclear weapons, the load being a variable mix of gravity bombs and air-to-surface missiles. The bombs can be in the megaton range (that is, equal to 1,000 kilotons) and can be delivered with very high accuracy.

(This accounting of the U.S. strategic



SELECTED ITEMS from the record military-budget request for fiscal year 1975, submitted recently to Congress by the Department of Defense, reveal a sharp across-the-board increase in the amount of money projected to initiate or expand programs that would increase U.S. "counterforce" capability against military targets in the U.S.S.R. (Colored bars are for the fiscal year 1975, gray bars for fiscal year 1974.) Items 1 through 6 refer to programs that will directly affect U.S. counterforce capability by improving the accuracy of land-based and sea-based missiles and by increasing the explosive yield of nuclear warheads. Items 7 through 11 refer to programs that involve indirectly related improvements in missile guidance, warhead multiplicity and warhead yield. Because this list was compiled from several different sources, there is some possibility that programs were double-counted or overlooked.

forces does not include the extensive U.S. "tactical" nuclear forces, many of which could attack targets in the U.S.S.R. In addition to the more than 7,000 tactical nuclear weapons in Europe, many such weapons are deployed in Asia and on forward-deployed ships in the Atlantic and the Pacific.)

In short, the U.S. already has a considerable potential for "limited" strategic strikes. Exactly how much capability depends on the critical assumption of who strikes first and how, as well as on one's assumptions about the nature of the Russian threat. In any case three important factors should be remembered about potential targets in the U.S.S.R.:

1. There are many nonmilitary, industrial targets outside urban centers that would require only one or two nuclear warheads each; such targets include manufacturing plants, power plants and the two construction yards for missile submarines.

2. Except for "hardened" targets, most military targets could be destroyed by only one or two warheads each; such targets include air-defense sites, military airfields, major army bases and submarine bases.

3. Even for hard targets such as missile silos, nuclear-weapons storage facilities and command posts, the use of small numbers of warheads will create a high probability of destruction. For instance, three Minuteman III warheads delivered against three Russian missile silos with a "hardness" about the same as that of the U.S. silos when they were first built would have approximately an 80 percent chance of destroying one silo, whereas seven Minuteman III warheads would have a similar 80 percent probability of knocking out one silo three times as hard. Presumably many Russian missile silos have a hardness in this range.

As a result, even with existing missiles a limited strike by the U.S. that employed 100 missiles or fewer could do substantial damage to the U.S.S.R. and could knock out some Russian ICBM's.

In calculating the sufficiency of our strategic forces, one should not forget the Chinese. For any conceivable "crisis scenario" the total expenditure of U.S. warheads against China could easily come from the present surplus exceeding the weapons needed for the assured-destruction mission against the U.S.S.R. Not only could the U.S. destroy most of the nascent Chinese nuclear forces, but also it has been estimated that a few warheads detonated over 50 Chinese urban centers would destroy half of the urban population (more than 50 million people), more than half of the industrial capacity and most of the key governmental, technical and managerial personnel. Indeed, against fixed targets such as cities the U.S. could use its B-52's, which could return to their bases for other missions.

Not only does the U.S. have this multifaceted capability but also its nuclear strategy has always included plans for attacks other than massive ones on Russian cities. This conclusion is logically inescapable when one realizes that the U.S. has had thousands of strategic warheads since the mid-1960's, has about 7,500 now and is expected to have almost 10,000 by 1977. There are only about 200 major cities in the U.S.S.R. Either the U.S. has aimed a superfluously large number of warheads at each major city or it has planned for other targets all along. Any doubts on this score were resolved by Secretary Schlesinger's statement in March that "our war plans have always included military targets."

President Nixon has made it very clear from the early days of his Administration that he wanted changes in U.S. strategic policy. Neither he nor any other high official, including Secretary Schlesinger, has ever rejected the assured-destruction concept. Rather they have defined assured destruction narrowly to mean only massive retaliation against cities and have said that more options are needed. To date the Nixon Administration has really presented two different sets of what "more" is needed. First there were the "sufficiency criteria," which were publicized in the period from 1970 to 1972. This past year has seen the emergence of a new set of criteria.

The sufficiency criteria, which President Nixon first hinted at in 1970, were spelled out by Secretary of Defense Melvin R. Laird in 1971. They are:

1. "Maintaining an adequate second-strike capability to deter an all-out surprise attack on our strategic forces."

2. "Providing no incentive for the Soviet Union to strike the United States first in a crisis."

3. "Preventing the Soviet Union from gaining the ability to cause considerably greater urban/industrial destruction than the United States could inflict on the Soviets in a nuclear war."

4. "Defending against damage from small attacks or accidental launches."

These four criteria have been explained further, including the fact that the deterrence is for the benefit of U.S. allies as well as the U.S.

The publication of the sufficiency criteria at least moved the public debate off the misleading view that U.S. policies and forces only envisioned massive retaliation against cities, but beyond that there is little new in the criteria. This is partly because they were never clearly explained; accordingly they remained more Delphic than definitive.

The first criterion is simply a basic statement of the assured-destruction concept. The third is a result of the assured-destruction assumption at meaningful levels of destruction; beyond the ability of either side to inflict 75 million fatalities and between 50 and 75 percent industrial damage—levels that would finish either country as a viable society—relative differences in the ability to inflict urban or industrial damage seem insignificant. Besides, much higher levels of destruction can only be achieved with considerable difficulty, since either country soon reaches a point of rapidly diminishing returns in terms of urban or industrial destruction per additional warhead.

The fourth criterion was clearly justification for the Safeguard ABM system. Without getting into the debate over such issues as whether or not the advantage of damage limitation against small attacks or accidental launches outweighs the disadvantage of the Russians' misinterpreting the purposes of any ABM deployment, suffice it to say that the Administration as early as May, 1971, was committed to insignificant ABM levels in the ongoing Strategic Arms Limitation Talks (SALT). The fourth criterion thus became "inoperative."

That leaves the second criterion. It clearly enunciates a desirable objective in strategic policy: to avoid strategic forces or actions that would be destabilizing in a crisis. Although this objective was not explicit before, it was inherent in the assured-destruction objective of providing highly survivable forces that would thereby reduce the incentive for a first strike. The second sufficiency criterion fails to delineate what more, if anything, was needed.

The criteria are silent about the kinds of option other than assured destruction that the President was so concerned about. Moreover, should the U.S. react to protect its allies (still undefined) in the same way that it would to protect its own territory? And what are U.S. strategic objectives with regard to China? In short, except for the flirtation with the ABM possibility, the sufficiency criteria only hinted at new strategic policies rather than establishing them.

Instead of trying to amend the sufficiency criteria, the Administration decided about a year ago simply to scrap them and to start anew in redefining strategic policies. This time Secretary Schlesinger has been the principal spokesman. After some of his press conferences late in 1973 and early in 1974 led to confusion among journalists and other observers as to what the new policies encompassed, the appearance of Schlesinger's annual report in March clarified the issues considerably. At one place in that report the "Principal Features of the Proposed Posture" (a posture Schlesinger clearly likes to refer to as "essential equivalence") are listed:

1. "a capability sufficiently large, diversified, and survivable so that it will provide us at all times with high confidence of riding out even a massive surprise attack and of penetrating enemy defenses, and with the ability to withhold an assured destruction reserve for an extended period of time."

2. "sufficient warning to ensure the survival of our heavy bombers together with the bomb alarm systems and command-control capabilities required by our National Command Authorities to direct the employment of the strategic forces in a controlled, selective, and restrained fashion."

3. "the forces to execute a wide range of options in response to potential actions by an enemy, including a capability for precise attacks on both soft and hard targets, while at the same time minimizing unintended collateral damage."

4. "the avoidance of any combination of forces that could be taken as an effort to acquire the ability to execute a first-strike disarming attack against the USSR."

5. "an offensive capability of such size and composition that all will perceive it as in overall balance with the strategic forces of any potential opponent."

6. "offensive and defensive capabilities and programs that conform with the provisions of current arms control agreements and at the same time facilitate the conclusion of more permanent treaties to control and, if possible, reduce the main nuclear arsenals."

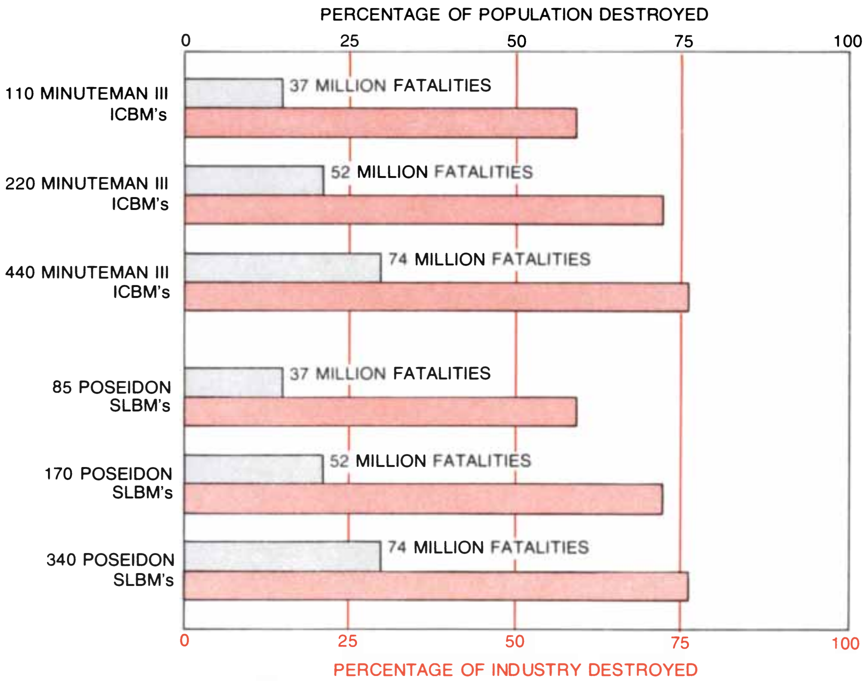
These factors plus the accompanying text in the report provide the best available insight into the proposed new policies. The first factor, combined with the second's requirement of bomber survivability, constitutes essentially a restatement of the assured-destruction concept. It needs no further elaboration here except to note that assured destruction does not require immediate response; indeed, the emphasis on a "second strike" capability and on the survivability of U.S. forces reflects the goal of having time in which to consider what the appropriate response should be.

Skipping briefly to the fourth, fifth and sixth factors, they raise a host of diverse issues—touching on all offensive and defensive strategic programs. There is not sufficient space to treat them comprehensively here; instead the focus will be on their impact on the Administration's concepts of strategic flexibility and limited nuclear war.

The third factor and the balance of the second address the questions of flexibility and limited strategic war directly. The underlying questions can best be summarized as follows: (1) Should the U.S. have a number of response options? (2) Should the U.S. develop missiles with improved counterforce capabilities? (3) Should the U.S. actively promote the idea of improving counterforce capabilities for fighting, if necessary, a limited nuclear war? Since the first question is essentially noncontroversial, the remaining two define the immediate issues.

Schlesinger reports that most of the targeting options in the past have involved "relatively massive responses." He wants to provide the President with a "wider set of much more selective targeting options." There is general agreement among strategic analysts that the U.S. should have a variety of response options other than massive retaliation against cities. These options could be useful, for example, in deterring a limited strategic attack. As Paul C. Warnke, a former Assistant Secretary of Defense, has put it: "There can... be little objection to the concept that our targeting plans should be sufficiently flexible to provide the President with a variety of options in the event of a nuclear attack." Warnke believes "we might be better positioned to deter a less than all-out Soviet attack if we have the refinement of command and control to push only one or a few buttons rather than the entire console... to respond with less than our Sunday punch."

This broad consensus includes those options that draw on the capabilities of present forces and those already well along in development. As we have seen,



COMPARATIVELY SMALL FRACTIONS of the strategic offensive forces of the U.S. can still destroy much of the population and industry of the U.S.S.R. The figures used to make this bar chart are derived from a table released in 1968 by Secretary of Defense Robert S. McNamara. That table was expressed in terms of alternative numbers of "one-megaton equivalent" warheads; this adaptation uses conservative assumptions to calculate the megaton equivalents of the multiple-warhead Minuteman III and Poseidon missiles. For example, only 10 warheads were assumed to be deployed on each Poseidon instead of the possible 14. The chart is based on the further assumption that the warheads are delivered; since missile reliability is not perfect, a few more missiles would have to be given the order to launch in each case to have the indicated number of warheads reach their targets.

our present forces already have the accuracy-yield combinations to be used effectively to destroy almost anything except hard targets. Even against such hard targets as ICBM silos these forces could destroy large numbers of targets, but they would not do it "efficiently."

Schlesinger makes it clear, however, that he wants more than flexibility, that he wants counterforce options that require new or improved weapons. The incremental options are ones "minimizing unintended collateral damage" and providing a hard-target kill capability that "matches" that of the Russians. To be able to achieve these options Schlesinger seeks programs to develop missiles with improved counterforce capabilities.

The proposed defense budget for the fiscal year 1975 includes a number of such programs. The programs appear to fall into two categories.

First, there are the short-term programs, the ones that involve relatively minor changes and for which initial deployment might easily begin by the late 1970's. The major programs in this category include procurement of more Minuteman III missiles; refinement of the existing guidance system of the Minuteman III to increase accuracy (probably from 1,500 feet down to 700 feet or less); a higher-yield warhead for the Minuteman III identical in configuration with the existing warhead, and a general program to improve and measure the accuracy of SLBM's. The proposed budget also includes funds to flight-test a Minuteman III with a larger number of smaller reentry vehicles. Whether this program will increase counterforce capabilities or not depends on the accuracy and yield of the new warheads.

Second, there are two major long-term programs. Both will require considerable development time, and initial deployment would seem unlikely before 1980. Advanced development will be initiated for a terminally guided "maneuverable reentry vehicle" (MARV) for possible "retrofit" into both ICBM's and SLBM's. Although a MARV warhead has been programmed for some time for the advanced Trident I SLBM, it is not to be terminally guided, being designed for evasion of ABM interceptors rather than for improved accuracy. A new terminally guided MARV, however, will presumably have an accuracy of a few hundred feet. This would give even warheads the size of the Poseidon's a very effective hard-target kill capability.

Further research and development is needed to decide exactly how the new MARV will work. By definition, after the MARV has separated from the "bus," or postboost vehicle, that holds all a missile's warheads, it can maneuver almost up to impact in order to correct its flight path. The corrections could be accomplished in two ways. The most likely development is the homing MARV, what some call the true MARV [*see illustration on page 28*]. A sensor in the warhead would acquire an image or images of the target or of prominent terrain features nearby (or perhaps would simply acquire an "altitude profile" of the terrain along its flight path). An on-board matching device would match this information with a map stored in its memory. The warhead's flight path would then be corrected either by gas jets or by aerodynamic vanes.

An alternative approach is to use an inertial guidance system in the warhead as well as in the bus. Since the reentry vehicle often separates from the bus early in its flight, an on-board guidance system would allow much later changes in trajectory. The information on position would come, however, from the system's gyroscopes, from stars or even from satellites and not from the target area itself. As a result this approach in theory would probably not be as accurate as the homing approach.

The second long-term program is the development of an entirely new ICBM for the 1980's. This missile, which may even be an air-mobile missile, would include a new guidance system (presumably a terminally guided MARV), which Schlesinger says would give it "a very good capability against hard targets."

How reasonable or necessary is it to develop missiles with improved counterforce capabilities in order to minimize collateral damage or to match the Russians' hard-target kill capability?

It is particularly difficult to understand how these missiles will minimize collateral damage. The warheads Secretary Schlesinger is proposing will probably have at least the yield of the present Minuteman III and Poseidon warheads. Such warheads would cause extensive damage over a wide area. For example, a "small" 100-kiloton bomb exploding in the air over a target would cause substantial fatalities and damage from immediate effects alone over a circle with a radius of 2.5 miles. Since the possible improvement in accuracy for the Minuteman, for example, is at most about 1,000 feet even in the long run, the number of civilian fatalities will hardly be reduced significantly if a warhead at least three to 11 times the size of the Hiroshima bomb lands a few hundred feet closer to the intended target.

A substantially smaller warhead that still provides an improved hard-target kill capability is unlikely to be ready for deployment until the 1980's, since a very accurate terminally guided MARV is needed to allow a significant "trade-off" between lower yield and higher accuracy. Furthermore, the value of much smaller warheads in saving lives must be put in perspective.

First, the way to minimize fatalities, if nuclear weapons must be used, is careful target selection, in other words aiming at targets distant from urban centers. Air-defense sites or air bases in the Arctic and isolated army posts or industrial sites are good examples. For a very limited exchange the differences in fatality levels from an attack on such targets with warheads of, say, 50 kilotons as against five kilotons would not be significant.

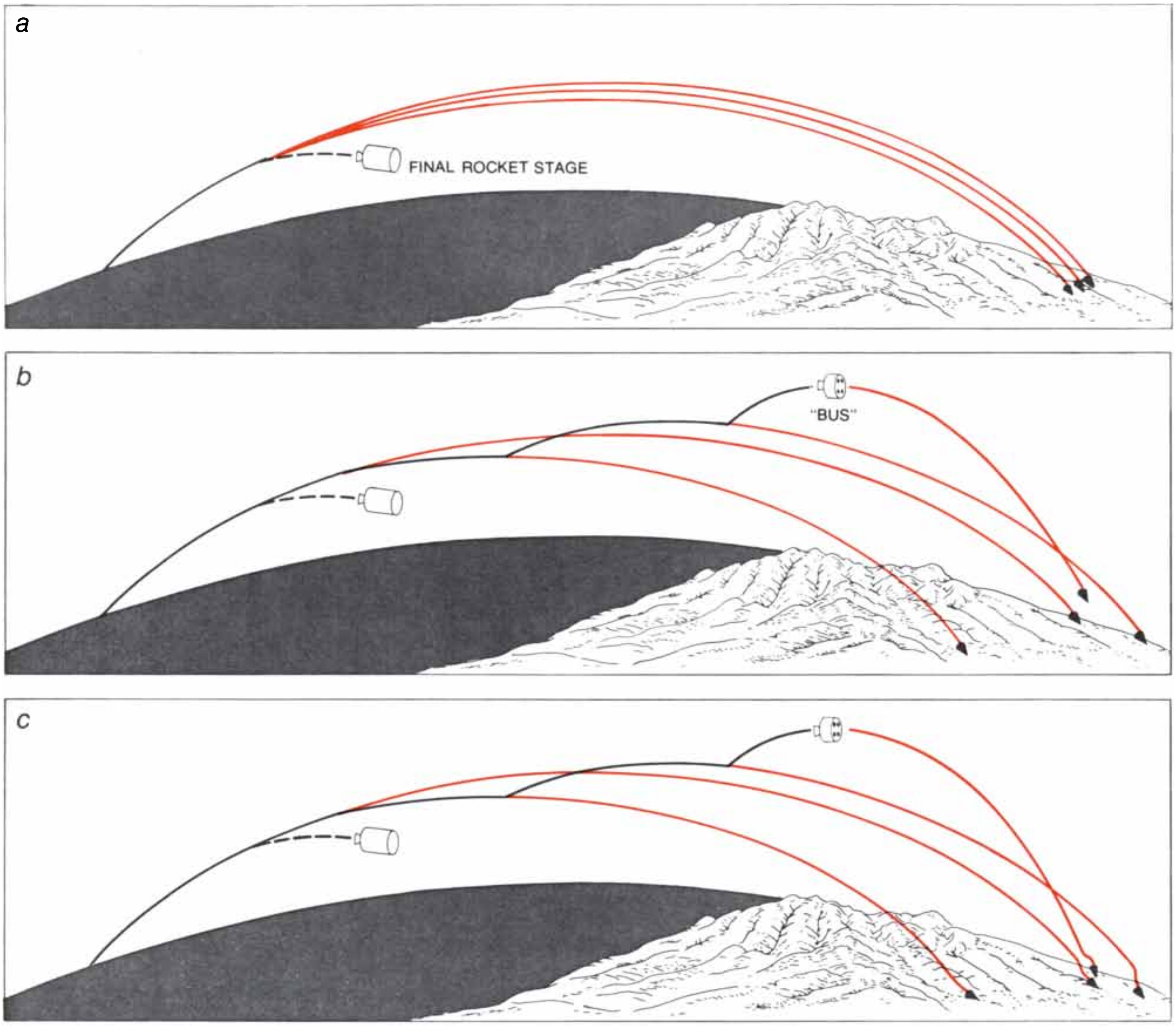
Second, if there is a large-scale nuclear exchange, then there simply is no way of keeping civilian damage at a low level. The effects not only of immediate blast but also of radioactivity would kill millions.

Third, in an actual nuclear exchange the successful continuation of a U.S. policy aimed at minimizing civilian casualties depends in large part on what the Russians do, and the Russians have never seemed much attracted to this objective. Their strategic warheads have always been large. Even though they necessarily reduced the size of individual warheads on their ICBM's in order to deploy MIRV's on them, some if not all of the warheads are still in the megaton range.

Schlesinger's main justification for the new counterforce programs is that the U.S. needs an "efficient" hard-target capability to match that of the U.S.S.R. This seems a questionable refinement of the broader theme of "essential equivalence." Schlesinger has on occasion defined essential equivalence to suggest overall balance. For example, he recently testified: "We do not have to have a match for everything in their arsenal. They do not have to have a match for everything in our arsenal."

Whether or not such an overall balance exists today and for the foreseeable future is a question that deserves public debate; a good case can be made for the affirmative. Most important, both the U.S. and the U.S.S.R. have a high-confidence ability to carry out a wide variety of retaliatory options. In terms of static indicators the Russians do have more missiles and greater missile "throw weight." The U.S., however, has more bombers, more warheads (now and for the rest of the decade) and about equal overall throw weight (if bombers are included in the calculations). In terms of qualitative factors U.S. missile submarines are much quieter and hence harder to find than the Russian ones, and U.S. bombers are more modern. Finally, to maintain or even enhance some of its capabilities, the U.S. already has a number of strategic programs well along: the conversion of older missiles to larger Minuteman III and Poseidon missiles, the B-1 bomber and the Trident submarine with its advanced missiles.

Schlesinger, however, avoids the complex question of whether the general U.S.-U.S.S.R. strategic picture is one of overall balance—of essential equivalence. Rather, he selectively focuses on relative counterforce capabilities against ICBM silos. Selective vision is not exactly a new tactic in military analysis. The "missile gap" of 1960 is a classic case; the heated debate over the number of U.S. ICBM's compared with the number of Russian ICBM's ignored the massive U.S. bomber force. Schlesinger's selective vision is even blurred within its own field. Although the Russians are clearly developing new missiles and MIRV's, they apparently have not pursued the accuracy aspect of a counterforce strategy with much zeal. As General George S. Brown, the chief of staff of the Air Force, recently remarked about the new Russian programs, "MIRVing alone won't [take out the Minuteman force]. Accuracy is the other



MULTIPLE-WARHEAD MISSILES come in several varieties. The U.S. Polaris A3 missile was equipped with multiple reentry vehicles, or MRV's; it released its warheads simultaneously at a comparatively early stage in the missile's flight; the separate warheads then continued, unpowered and unguided, along the missile's initial ballistic trajectory, falling in a tight pattern (*a*). The more advanced Minuteman III and Poseidon missiles, now being deployed by the U.S., are equipped with MIRV's (*b*). These warheads are carried on a "bus," which is capable of changing its direction in flight; after each such change the bus can release another warhead, which then continues, again unpowered and unguided, along that ballistic trajectory. The warheads can fall on widely separated targets. (The Russians started testing MIRV's on some of their missiles last summer, and it is estimated by the U.S. Department of Defense that they could begin deploying them in 1975.) The U.S. now proposes to develop maneuverable reentry vehicles, or MARV's, with terminal guidance (*c*). Each of these warheads will be capable of greater accuracy (*see illustration on page 28*).

key element and we haven't seen evidence of accuracy improvement in their work which we would expect to see."

Is there some reason why the U.S. and the U.S.S.R. should have essential equivalence in the capability to destroy missile silos? The arguments against this course of action seem persuasive. There is no benefit in terms of traditional strategic analysis in being able to kill efficiently very large numbers of the other side's silos. As we have established, the U.S. can already destroy some silos, although at a cost of a few U.S. missiles each. Inefficient, limited destruction of silos should suffice for the war scenarios that some envision, in which the U.S. feels it necessary to destroy silos as a way of showing its "resolve." Killing many more silos would not minimize damage to the U.S.; everyone agrees that the U.S. cannot expect to destroy a large enough fraction of the silos or other strategic offensive forces of the U.S.S.R. to limit damage to this country in any meaningful way.

Finally, a critical assumption underlying the preceding discussion is that the silos will have missiles in them when they are destroyed. In fact, the flight time of a Minuteman missile to the Russian missile fields is about 30 minutes. If the Russians were to deploy early-warning satellites, they could detect almost instantaneously the launch of U.S. missiles, which means that the U.S.S.R. could probably have the option of launching many, if not all, of its missiles before the U.S. warheads arrived. Using U.S. warheads against empty silos in empty fields seems a particularly questionable policy.

The full cost of these new programs is unclear. Much depends on the size of the deployments and the extensiveness of the modifications. A useful benchmark is the Minuteman III program; the conversion of 550 older Minuteman missiles into Minuteman III's will cost between $5 billion and $6 billion. Although the costs of some of the new counterforce programs might be comparatively small, the total cost of all the new programs would greatly exceed the Minuteman III costs.

Added to the questions about the analytical reason for the new counterforce programs and the inevitable costs must be the distinct possibility that these programs will be destabilizing and will make arms limitations more difficult to negotiate.

Assuming a crisis situation, a substantial U.S. counterforce capability against Russian ICBM's is more likely to create an incentive for the U.S.S.R. to adopt a hair-trigger, launch-on-warning posture; the Russian leadership would fear that the U.S. might attack first in an attempt to limit damage to itself. These fears would make it even more likely for the U.S.S.R. to attack first in a crisis in order to destroy some of the U.S. ICBM's that had become more tempting targets as a result of the new U.S. counterforce programs.

Schlesinger deplores this instability (as in his fourth feature, cited above, of the new posture), but he and other high officials say that the new U.S. programs are not extensive enough to create such Russian fears. The conceivable accuracy and yield improvements on 1,000 Minuteman missiles, however, even without the terminally guided MARV, could give the U.S. the capability, on paper at least, of destroying between 80 and 90 percent of the Russian ICBM force. The deployment of the MARV or the use of improved SLBM's against the Russian missiles would push that percentage even higher.

The Russian leadership, moreover, might be more conservative than the U.S. leadership in assessing Russian strengths and weaknesses. This conservatism would be based at least partly on the fact that, unlike the balanced reliance in the U.S. on all three elements of the strategic triad, in the U.S.S.R. ICBM's are the primary component of the strategic offensive forces. The U.S.S.R. is allowed up to 1,618 ICBM's under the SALT I Interim Agreement (compared with 1,054 for the U.S.), and the Russians are actively developing four new ICBM's. Moreover, these missiles are under the command of the Strategic Rocket Forces, which since it was created in about 1960 has been one of the most important branches, if not the most important one, of the Russian military. Unlike the U.S. Air Force, which has responsibility not only for ICBM's but also for bombers and many tactical forces, the primary responsibility of the Strategic Rocket Forces is the Russian ICBM force; consequently this organization has every incentive to enhance its role in strategic planning. The Long Range Aviation command, which has responsibility for the Russian bombers, has never had the bureaucratic strength of the Strategic Rocket Forces, and the Russian navy has responsibility for a number of other forces besides missile submarines.

The strategic-planning emphases of the U.S. and the U.S.S.R. differ particularly on the subject of bombers. At present the U.S. has more than 450 intercontinental bombers, about a fourth of which are kept on "ready alert" at a large number of air bases (so that they can avoid being destroyed even in case of surprise attack). The Russians have about 140 long-range bombers. These are qualitatively inferior even to the B-36 bombers deployed by the U.S. in the 1950's, are not kept at as high readiness and are located at just a few air bases. Although a new Russian bomber (named the Backfire by the Pentagon) is just beginning production, it seems primarily intended for targets on the periphery of the U.S.S.R. In any case it is not certain how many Backfires will be built, and the plane appears to lack the critical range and low-altitude capabilities of the B-52's.

As for SLBM's, the U.S.S.R. is building new missile submarines and is allowed more boats and SLBM's than the U.S. under the terms of the SALT agreements. In contrast to the active U.S. MIRV programs for both ICBM's and SLBM's and the new Russian MIRV programs for ICBM's, however, the Russians have not begun testing multiple warheads on their new SLBM. The U.S.S.R., moreover, usually keeps only five or six missile submarines on patrol at any one time, compared with 40 percent of the 41 U.S. boats. In sum, the U.S.S.R. does not seem to give missile submarines the same priority in strategic planning as the U.S.

Schlesinger essentially hinges his denial that first-strike fears by the U.S.S.R. would be enhanced by the planned U.S. improvement in its capabilities against ICBM's on the relative invulnerability of the Russian missile submarines. Compared with the U.S. missile submarines, however, the Russian boats are noisier—an important qualitative disadvantage—and must operate in ocean areas where it is easier for the U.S. to locate and detect them. In addition the U.S. has under way a large, aggressive antisubmarine-warfare program for tactical and strategic uses. It has been reliably estimated that U.S. expenditures in the fiscal year 1972 for antisubmarine warfare were $2.5 billion and that by 1974 they would rise to more than $4 billion. The Russian leaders might well fear, at some future crisis point, that the U.S. had developed a significant antisubmarine-warfare capability, making Schlesinger's suggested ultimate reliance on their missile submarines less than completely reassuring.

One "crisis scenario" that is often concocted to show the danger of the growing Russian counterforce capability against Minuteman and to justify developing improved U.S. counterforce capabilities is an attack or threat of attack by the U.S.S.R. against U.S. ICBM's. The scenario envisions the following chain of events: (1) a real or threatened Russian attack against Minuteman: (2) a realization by the U.S. leadership that it is left or will be left with no more than a capacity to attack Russian cities: (3) major concessions or even surrender by the U.S.

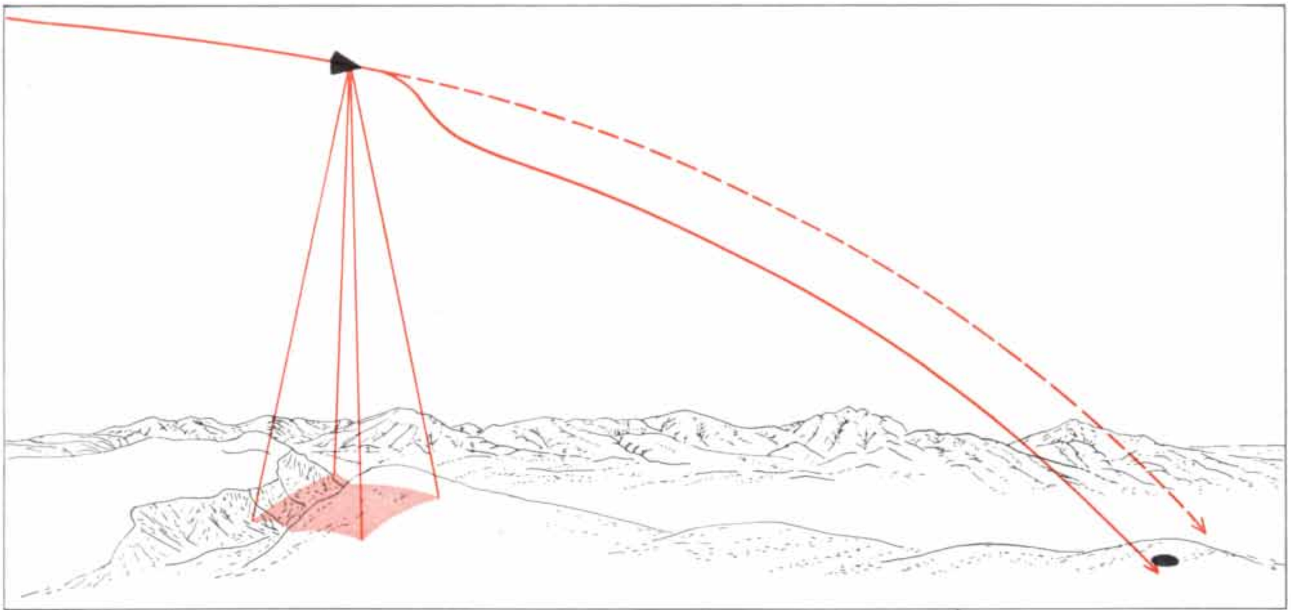
This scenario has an obviously fantastic quality. Even if the internal logic of the scenario were accepted, it still does not justify improving U.S. counterforce capabilities. It does not matter whether the U.S. missiles destroyed are highly accurate or not. What matters is what other U.S. forces can do if these missiles are destroyed. Indeed, as we have seen, by presenting an increased threat to the U.S.S.R., U.S. development of highly accurate missiles might actually make the Russians more likely to attack, thus making the scenario less implausible.

More important, the underlying logic of the scenario is simply wrong, as should be evident to both the U.S. and the Russian leadership. First, the Russians would have to consider that Minuteman might be launched against Russian targets in the 30-minute warning time between the launch of the Russian ICBM's and their arrival at the Minuteman silos. Second, even if a surprised or reasonably cautious U.S. leadership did not launch on warning, a few Minutemen would survive even the most careful attack. Also surviving would be at least the bombers on alert and most if not all of the U.S. missile submarines in the water. (If the attack occurred after an initial crisis period, more bombers than usual would be on alert and more submarines would be in the water.) These combined forces would provide the U.S. with the capacity to carry out a number of limited strikes while still retaining an assured-destruction hedge.

Finally, some U.S. retaliation would seem very likely to the Russian leadership since tens of millions of Americans would be killed in any "Minuteman only" attack. In attacks against silos the bombs are set to explode as close to the ground as possible, thereby picking up much dirt and debris. The fallout from the explosion of thousands of megatons of nuclear weapons over the Minuteman fields would be tremendous, and winds would carry the lethal contamination over many major U.S. cities [*see illustration on page 30*]. Such calculations of fallout do not even include the possibility of a few Russian warheads going off course and directly hitting populated areas, nor the collateral damage by Russian attacks against other targets, such as bomber bases, many of which are near cities [*see illustration on page 31*].

Even not assuming a crisis, the consequence of these new U.S. counterforce developments might be to push the U.S.S.R. toward accelerating or expanding existing programs, or starting new ones. The arms race is not as mechanically "action-reaction" as some have suggested, but a substantial new U.S. capability against the primary strategic offensive force of the U.S.S.R. will surely fuel justifications within the Russian bureaucracy for some kind of reaction. This should be particularly true when U.S. antisubmarine-warfare programs, noted above, are also considered.

If the U.S. counterforce programs are allowed to continue beyond the rhetoric of announcing them, these programs would operate to undercut any progress at SALT. Of course, if announcing these programs is just a short-term ploy designed to strengthen the U.S. bargaining position for the impending SALT II



MOST LIKELY CANDIDATE for the proposed terminally guided MARV is the "homing" MARV, depicted in this idealized drawing. (The homing MARV is sometimes referred to as a "true" MARV to distinguish it from another type of terminally guided MARV that would rely on a comparatively straightforward, on-board inertial-guidance system.) The homing MARV would operate by means of an on-board sensor capable of acquiring an image or images of the target area or of some nearby, prominent terrain feature. (Alternatively, the sensor could operate by obtaining an "altitude profile" of the terrain features along its flight path.) It is still unclear what kind of sensor will be used; radar and optical systems are considered among the leading possibilities. Once the target or other feature has been acquired the warhead would be maneuvered under the control of an on-board computer on the basis of feedback information from the sensor. The actual terminal maneuvering could be accomplished by means of gas jets or aerodynamic vanes.

agreements, then little real harm will result. There is no evidence, however, that top Administration officials intend to turn these programs off quickly. And even if there are such intentions, new weapons programs tend to gain a momentum of their own once they are announced. High-level officials become publicly committed to rationales for them, rationales that include more than the systems' just being "bargaining chips." Bureaucracies are created with a vested interest in the continuation and expansion of these programs. Moreover, improvements in accuracy and yield would be particularly difficult to limit explicitly in SALT, making it harder to rationalize publicly any subsequent termination of the program.

Accuracy improvements are generally accepted as being among the most difficult weapons characteristics to limit in an arms-control agreement, because of problems of both definition and verification. Drafting a workable, direct limit on accuracy seems impossible, since the counterforce potential of a warhead depends on the accuracy-yield combination. Moreover, a simple numerical limit on accuracy would not be verifiable. A photograph of a silo or even the missile gives little clue to the kind of small but important differences in accuracy that are being considered here. Closer examination through on-site inspection, even if such inspection could be negotiated, would be insufficient. On-site inspection could indicate whether the warhead was a terminally guided MARV, but this would not establish any particular accuracy. Moreover, on-site inspection includes a heroic assumption that the latest warheads are on the missile and not stored nearby in an area excluded from the on-site inspection provisions.

Surveillance of Russian missile-testing may give some indication of accuracy. The indication, however, is indirect and not conclusive. Test data tell one about the ballistic coefficient (or pointedness) of the warhead, its reentry speed and similar information, all of which helps in estimating accuracy. An outside observer, however, can never be sure what the actual target is. Similarly, course corrections by the warhead would indicate a maneuvering capability but not necessarily terminal guidance or particularly high accuracies.

An indirect way to limit or impede accuracy improvements through SALT would be by placing a strict limit on the number of missile tests. This would make it more difficult to develop advanced guidance techniques and to test them often enough so that the military would have confidence in them. The low limits necessary seem nonnegotiable, however, since they represent a direct challenge to all new strategic programs. Even without accuracy improvements the Pentagon will want to do extensive research and development and operational testing of the new Trident missile and further operational testing of the Minuteman and Poseidon missiles. Similarly, the Russians will want to flight-test extensively their four new ICBM's and their new SLBM as well as their existing arsenal of missiles.

Limits in SALT on the yield of warheads might be more possible, but they would be of uncertain significance. The two sides could limit yield by an agreement that warheads not be larger than a given yield or a given weight. The effect of any such limitation could be circumvented, however, by increasing the number of warheads and by increasing their accuracy. Moreover, it would be difficult to verify the exact yield of a warhead. Even elaborate on-site inspection would not ensure that "advanced" warheads were not hidden nearby. Surveillance of flight tests only gives an estimate of the size of the warhead, and yield per pound of warhead can be varied by warhead design and the richness of the nuclear "fuel" used.

In short, the practical difficulties of fashioning limitations in SALT on the type of counterforce improvements now planned by the U.S. make such limitations unlikely and will instead presumably create strong pressures in the U.S.S.R. to expand old programs or to start new ones that either match or compensate for the U.S. programs. This in turn can only work against other limitations on strategic arms.

Allied concerns about the credibility of the U.S. deterrent are another reason offered for developing missiles with improved counterforce capabilities. Occasionally a specific scenario—a Russian attack in central Europe—is given as a justification for such improvements. Neither the scenario nor the more general invocation of allied claims is persuasive.

The European scenario supposedly demonstrates that the U.S. needs the ability to respond with nuclear weapons in order to show its resolve and to destroy some of the attacking Russian forces. There are, however, already sizable U.S. forces in Europe that could accomplish both of those objectives. Even if the U.S. decided to employ strategic weapons, existing U.S. forces could carry out a wide variety of selective attacks.

As for the broader claims of allied concerns, Morton Halperin, an authority on nuclear strategy, has remarked: "The credibility of the U.S. deterrent to an Ally is primarily a result of the overall U.S.-Ally relationship, which includes economic and political considerations as well as military. To the extent that Allied leaders evaluate U.S. military capabilities, they look especially to the U.S. conventional and nuclear forces in that particular theater of operations. Fine distinctions in the U.S.-Soviet strategic balance or in U.S. strategic policy are unimportant to Allied leaders. Among those Allied analysts who care, opinion is probably split between those who favor the U.S. possessing an efficient silo-kill capability and those who do not."

Among the European strategic analysts who oppose such deployments is Ian Smart, formerly assistant director of the London-based International Institute of Strategic Studies. Smart writes: "Producing and deploying much more accurate strategic missiles ... is to be regretted and even feared since ... it can only reduce the stability of the strategic balance in any period of acute tension." At least part of this European concern can be attributed to the fact that, in a strategic exchange, the industrialized European countries are very likely targets—if only because of the U.S. forces deployed in or near those countries.

Finally, even assuming that the allies (or even the American people) accord considerable political significance to fine distinctions in the "strategic balance," Schlesinger's proposed counterforce improvements are not very helpful politics. The supposedly important distinctions are usually visible ones such as the number of delivery vehicles, the number of warheads or the throw weight. Schlesinger's accuracy and yield improvements do not affect these indicators, except possibly in the counterproductive way of reducing the number of warheads in order to allow larger ones.

On balance, then, there seem to be strong arguments against developing missiles with improved counterforce capabilities. Collateral damage can best be minimized by shifting targets, not improving accuracies by a few hundred feet. The ability to destroy efficiently large numbers of missile silos in order to "match the Russians" seems not only unnecessary and expensive but also destabilizing. SALT might well be undercut, and the supposed concerns of our allies about the U.S. deterrent are not answered by such programs.

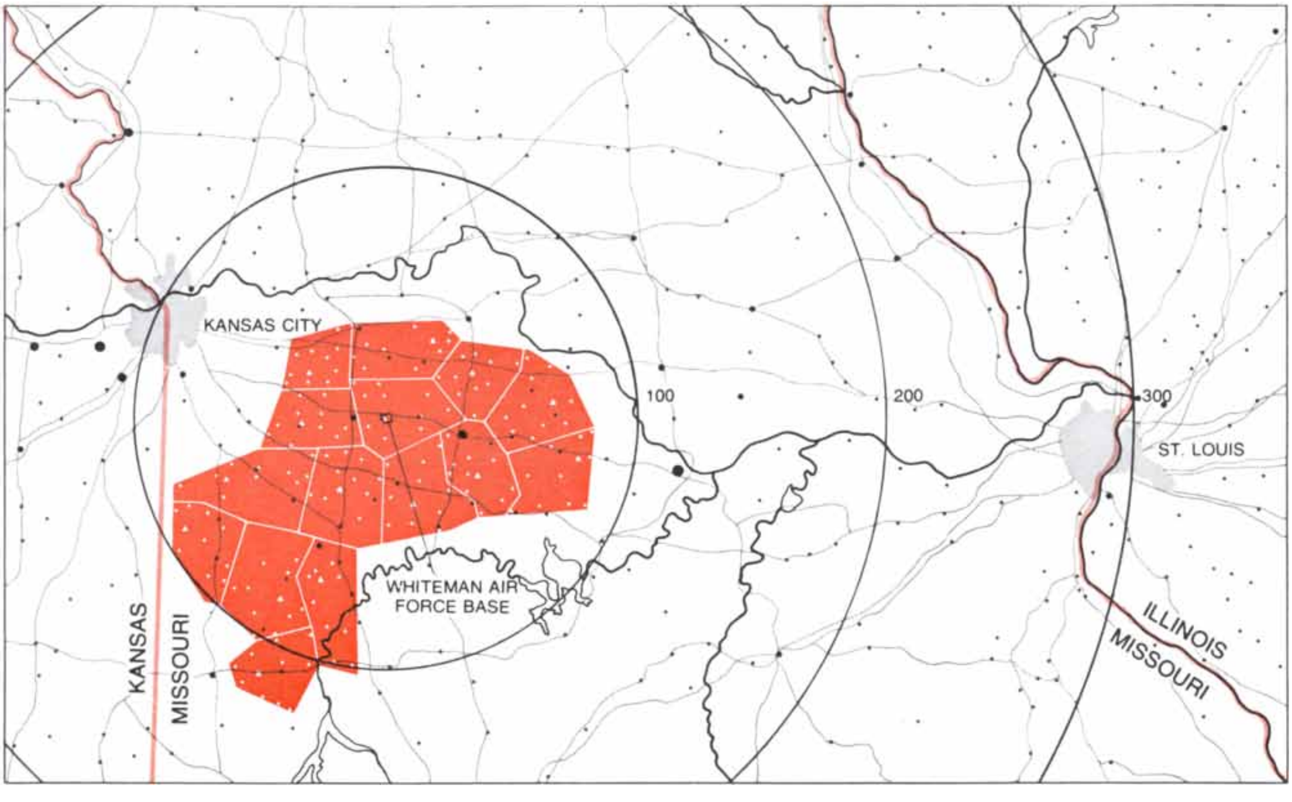
As one gets caught up in considering nuclear-war scenarios and nuclear-weapons capabilities there is a dangerous tendency to forget that the primary objective of nuclear strategy is to avoid nuclear wars, not to fight them.

Given the destructive power of nuclear weapons and the world's lack of experience in using them, crossing the "nuclear threshold" would be a profoundly destabilizing event. It is a delusion to believe one country could employ nuclear weapons, even on a limited scale, and have a high degree of confidence that the response by another nuclear power would be predictable and proportionate. The particular first use might be estimated by the opposing country's observers to be greater than it actually was, or the use might have created more damage than expected (for example through greater-than-expected fallout). The opposing country might not have readily available weapons of the same yield or similar targeting options and decide to escalate. The political reaction in the opposing country might lead to escalation. In short, the possible causes for matters getting out of hand are endless.

To make deterrence work, a country must carefully consider its public attitude toward nuclear war and cautiously select its retaliatory options. This does not mean that the U.S. should have only the single strategic option of massive retaliation against cities. This country already has ample capabilities for lesser options, and it seems appropriate to have the flexibility, at a minimum, for possible responses to accidental or limited launches.

The Nixon Administration, however, is going beyond this. It is seeking the additional capability to attack efficiently large numbers of Russian missile silos. Not only might this counterforce option be destabilizing in itself but also the Administration's promotion of the option and its general public advocacy of a counterforce strategy might have a pervasive, if subtle, tendency to reduce the inhibitions against the use of nuclear weapons—in effect, to lower the "nuclear threshold." New bureaucracies, with vested interests in the hardware and rationales of a counterforce strategy, are created. In trying to gain public approval of new policies and programs, leaders find themselves taking more simplistic positions than the uncertainty of nuclear warfare warrants. In this climate some of the risks of nuclear war are downplayed. Unrealistically precise calculations suggest that limited nuclear war can be kept limited and even result in positive gains.

There are some disturbing parallels here to the vogue of limited conventional



RUSSIAN PLANNERS MUST CONSIDER that a hypothetical "limited counterforce" attack against the U.S. Minuteman ICBM fields (such as the one shown here in the vicinity of Whiteman Air Force Base in Missouri) would create an enormous cloud of lethal radioactive fallout that would presumably extend for hundreds of miles downwind of the target area, killing or injuring millions of exposed civilians in nearby cities and towns and thereby making massive U.S. retaliation seem very likely. (For comparison, a single 15-megaton nuclear explosion at Eniwetok in 1954 contaminated an area of some 7,000 square miles, extending 300 miles downwind and 20 miles upwind; to be at all effective, any contemplated counterforce attack against a Minuteman field would require the ground-level explosion of hundreds of megatons of nuclear warheads and would accordingly produce a much larger area of radioactive contamination.) The Minuteman field at Whiteman, one of six such fields in the north-central U.S., has 150 Minuteman missiles arranged in 15 "flights" of 10 missiles each (*irregular colored shapes*); the missiles are installed in hardened underground "launch facilities," or "silos" (*white dots*), with each flight of missiles having its own "launch control facility" (*white triangles*). The concentric scale rings are centered on the main airfield at Whiteman and are drawn at 100-mile radial intervals. Gray lines are principal roads. Black lines are major rivers. Gray areas are main metropolitan regions. Black dots are smaller cities and towns.

war in the early 1960's. In pushing for changes in conventional strategy and new procurement, advocates of limited conventional war ignored some of the pitfalls and costs of such a strategy. The searing national experience of the war in Vietnam was needed to demonstrate these oversights.

Exactly where the line should be drawn on "selective targeting options" is not at all clear. It seems most inadvisable, however, to take the gamble of developing missiles with improved counterforce capabilities, whether this is to match a specific Russian capability or for any other reason.

Opponents of U.S. counterforce improvements, nonetheless, must recognize certain practical limits to their arguments. Even if Congress declines to fund the new and accelerated development programs Schlesinger is proposing, continued U.S. testing of strategic missiles and various research-and-development efforts already under way inevitably will lead to some improvements in missile accuracy. (As Schlesinger has pointed out, some refinements in existing guidance systems will occur almost as a matter of course—through better software programs, greater purity in rocket fuel, better measurement of the earth's gravitational field and numerous other factors. The development of a terminally guided MARV, something further beyond the state of the art, requires more of a conscious bureaucratic decision to proceed.) Besides U.S. advances, moreover, Russian counterforce improvements are likely to continue, raising serious questions about Russian intentions.

"UNINTENDED COLLATERAL DAMAGE" resulting from a "limited counterforce attack" against a military target such as a large airfield would also be considerable, particularly since many such targets are located near major cities. For example, Logan airport near Boston, shown on this map, must be considered a military target, since strategic bombers are likely to use it in a war. (B-52's were reportedly stationed there during the Cuban missile crisis.) A 100-kiloton nuclear warhead exploding directly over Logan would create an "overpressure" of between three and four pounds per square inch out to a radius of about 2½ miles (*smaller colored circle*); a one-megaton warhead would create the same overpressure out to five miles (*larger colored circle*). This overpressure would kill many people in the area and would cause severe to moderate damage to most structures. (For a surface explosion the damage radius would be about three-quarters of that for an airburst of the same yield, but the fallout would be greater.) An improvement in missile accuracy of a few hundred feet or so would clearly not alter the position of these large circles by much.

Faced with these likely developments, the solution is still not to follow the Schlesinger approach. Rather, the solution should be to seek actively to negotiate for limits on MIRV's and for the reduction of vulnerable strategic forces.

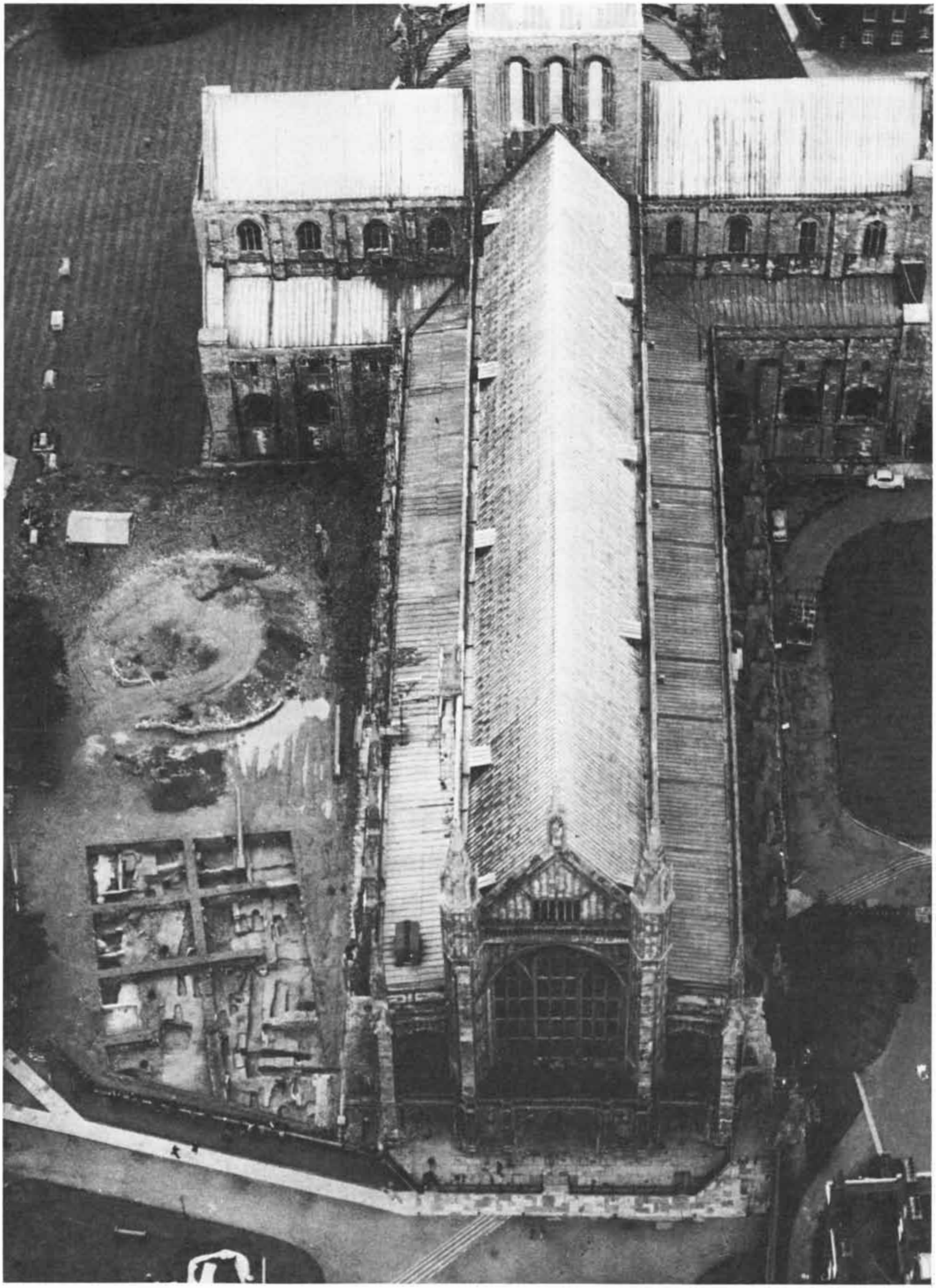
Limits on MIRV's would be designed to slow the perceived threat to U.S. ICBM's, a Russian threat that many consider destabilizing. In return for the U.S. slowing certain of its strategic programs, for example, the U.S.S.R. might agree to limits on the deployment of the SSX-18, the "follow on" missile to the large SS-9. This would push at least a few years further into the future the time when analysts would estimate that only a particular level of Minuteman could survive a Russian counterforce attack.

Negotiating missile reductions represents another approach: to limit not only the threatening forces but also the threatened ones. This approach would essentially mean bilateral reductions in ICBM's, presumably in a way that would retire the more threatening ICBM's, so that the remaining ICBM's would be less vulnerable. Some asymmetrical reductions might also be considered. For instance, the U.S. could reduce its ICBM's, whereas the U.S.S.R. (having less to fear in the short run about the vulnerability of its ICBM's) could reduce some ICBM's plus other forces.

Reductions in the land-based missiles of both sides would reduce the importance of this strategic strike force. It would thereby undercut the rationale for an expensive contest of matching counterforce improvements. More important, it would reduce the greatest potential source of instability in a crisis. Both countries would have less incentive to adopt an unstable, launch-on-warning posture or to launch an attack out of fear of a preemptive strike.

The reductions approach has received support recently from such diverse sources as the Federation of American Scientists and Fred C. Iklé, director of the Arms Control and Disarmament Agency. It was even accorded the status of a possibility in Schlesinger's recent annual report.

Rather than focusing on how to match the U.S.S.R. in a particular capability when such matching does not bode well for either country, the strategic debate in the U.S. in the coming months should focus on MIRV limits, force reductions and other measures designed to minimize the chances of nuclear war and to decelerate the arms race.



SITE OF OLD MINSTER, the principal cathedral church of the Anglo-Saxon kingdom of Wessex, is seen under excavation (*left*) in this aerial photograph. The excavation lies to the north of the cathedral built by the Normans following their conquest of England.

The Archaeology of Winchester

This English cathedral city was faced with the loss of its past as a result of urban redevelopment. Excavation has now revealed the pattern of its growth since its birth some 2,000 years ago

by Martin Biddle

How is a city born and how does it grow? If it is long dead, like Troy or royal Ur, archaeology can readily provide some of the answers. If the city is still very much alive, like Rome or London, the evidence is harder to obtain. Nonetheless, over the past 12 years an intensive archaeological campaign has uncovered the early periods and amplified the recorded history of one such city. The site was a major defended settlement during the latter part of the British Iron Age, was the island's fifth-largest town in Roman times, was a prosperous bishopric from the seventh century, was a royal seat until well after the Norman Conquest and is today one of England's leading cathedral cities, with a population of 33,000. The city is Winchester, and what more than a decade of urban archaeology has revealed about it is a fair indication of how much can be achieved elsewhere in the world when the work is begun before the past is irretrievably destroyed.

The River Itchen, the trout stream made famous by Izaak Walton in *The Compleat Angler,* rises in central Hampshire and flows south through a range of chalk downs on its way to Southampton Water, behind the Isle of Wight. Since remote antiquity the river valley and the grassy downs have provided natural lines of communication, the one north-south and the other east-west. At the point where the two routes cross and the alluvial valley floor is narrowest a spur of the chalk downs slopes more gently than elsewhere toward the riverbank.

A mile or so southeast of this spur the valley of the Itchen is commanded on its opposite eastern side by an Iron Age hill fort: St. Catharine's Hill [*see illustration on page 35*]. Built during the third or second century B.C. and enclosing an area of more than 20 acres, the defenses give evidence of having been reconstructed several times before being burned in the first century B.C. By then a settlement had appeared on the western side of the valley, on the same chalk spur where Winchester would later lie. In about the middle of the first century B.C. the new settlement was formally defined by the construction of a rampart and ditch that enclosed an area of just over 40 acres, or nearly twice the area of the eastern hill fort. This enclosure was the dominant feature of the valley in the later Iron Age. It lay astride the east-west route and commanded the river crossing.

Although little is known of the interior of the western settlement, its central area seems to have been densely occupied. The economy of the inhabitants was based on agriculture, but there is also evidence for long-distance trade connections. Fragments of southern Italian wine amphoras of the first century B.C. have been found, and the enclosure and its immediate vicinity have produced nine large bronze Ptolemaic coins of the third century B.C. from Egypt. These, of course, may have been imported for the value of their metal long after the time when they were first minted.

The size of the western enclosure is impressive, but there is not yet enough evidence to suggest that it contained an urban or even a proto-urban community. Moreover, the settlement appears to have been a false start; there is a break of as much as 100 years in its habitation. When the area was reoccupied soon after the Roman conquest of Britain in A.D. 43, the settlement was on the valley floor near the river, outside and downhill from the Iron Age enclosure.

Much of what we know about the growth of the settlement in Roman times is the result of emergency excavations first undertaken in 1961, when preparations were being made for the building of a new hotel in the center of the city. The rescue excavation soon revealed two facts. The first was the immense and virtually untapped wealth of Winchester's archaeological record. The second was the rate at which that record would be destroyed by the modern developments then planned for the decade ahead.

The city today is an important administrative, judicial, military, ecclesiastical and business center. The pressures generated by these urban functions find expression in plans for new roads and buildings, all potentially destructive of the buried remains of the city's past. A special body, the Winchester Excavations Committee, was set up in 1962 to deal with the problem of investigating and recording the archaeological evidence before its destruction. During the next 10 years the committee administered the largest program of urban excavation yet undertaken in Britain (or elsewhere in Europe for that matter). For seven years the work has been a joint Anglo-American venture, done in collaboration with the University of North Carolina and Duke University and supported by government and foundation funds from both sides of the Atlantic.

The project had from the start the principal objective of studying the origin and changing character of the urban community throughout its entire existence, from the first permanent settlement down to the emergence of the modern city in the reign of Victoria. The city itself was to be the subject, rather than any one period or aspect of its past. We hoped to try to grasp the totality of the urban phenomenon and the interaction of the city and its setting, both at distinct moments in time and between one period of its development and another.

The project involved not only rescue

excavations on threatened sites but also excavations on unthreatened ones, some of them large enterprises that yielded information essential to any balanced concept of the city's evolution. The project also required the integrated utilization of all the available evidence, whether it was from archaeology, from the natural sciences or from written records (in which the city is immensely rich from the 12th century on).

In 1968 the Winchester Research Unit was set up to prepare this large body of material for publication. A series of perhaps 12 volumes of *Winchester Studies* is planned. They will come from the Clarendon Press at Oxford and the University of North Carolina Press. The first volume will appear late this year or early in 1975.

The new Romano-British settlement on the Itchen was peopled not by Romans from Italy but by Romanized Celts. Their settlement may have grown up in a rather formless way at the junction of the new Roman roads that met close to the old river crossing. These roads, built shortly after the Roman conquest, were perhaps protected by a detachment of troops housed in a fort at or near their junction. If the fort ever existed, and the evidence is still unclear, its life was no more than 20 years. The development of the civil settlement was in contrast rapid: by the end of the first century it had become a walled city with a chessboard street plan and public buildings. The earth-and-timber ramparts of the city defenses enclosed an area of more than 143 acres. Their line was followed by all subsequent walls of the city down to the end of the Middle Ages.

The size of the enclosed area made Winchester the fifth-largest city in Roman Britain. Two long-distance Roman roads formed the axes of the rectilinear street system that divided the city into *insulae,* or blocks. A central block was occupied by the forum and basilica, constructed by about A.D. 100 to house the judicial, administrative and principal commercial functions of the city. The city was now known as Venta Belgarum. As "Venta" may imply, it was the market center of its region, and as "Belgarum" indicates, that region was populated by the Celtic tribesmen known as the Belgae, from among whose principal landowners the city's chief citizens were drawn.

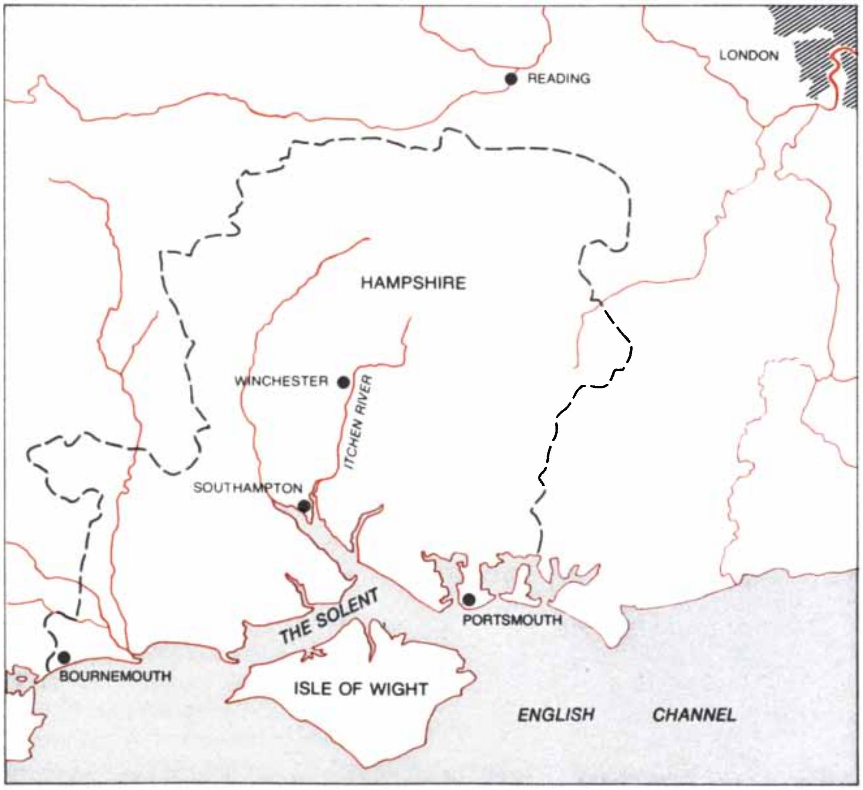
Little is known of the detailed development of Roman Winchester. At first its houses, even if they were provided with such amenities as glazed windows, painted walls, tiled roofs and mosaic floors, were built of timber. Increasing prosperity in the later second century led to their being rebuilt in stone. About A.D. 200 the defenses, which must long have been out of repair, were totally remodeled. Within a generation the earth-and-timber perimeter was strengthened by the addition of a stone wall that was to stand for 1,500 years. The influence of the stone ramparts is still reflected in the traffic problems of the modern city.

The Romano-Britons left behind one common kind of archaeological evidence: the remains of their dead. In the Roman fashion the cemeteries lay along the roads outside the city gates. Here in the mid-fourth century, among the burials of the native population, is found the first clear evidence of the arrival of aliens at Winchester. Some graves, distinguished by the leather belts with bronze fittings they contained, are apparently those of soldiers. Their equipment is of a kind well known along Rome's frontiers on the Rhine and the Danube.

During the second half of the fourth century, following the barbarian ravaging of Britain in A.D. 367, the defenses of the island provinces were reorganized. An important element in the revised strategy was a system of "defense in depth," which was to be provided by walled towns. The towns' fortifications were strengthened by the addition of projecting towers, evidently to mount catapults that could rake attackers with arcs of intersecting fire. Winchester was included in the defense system. The maintenance and use of such artillery probably required the services of specialized troops; the alien element now recognized in burials of this period at Winchester and at other towns may well indicate the presence of such specialists. They were foreigners (perhaps Germans), either regular soldiers or mercenaries.

The movement of troops and even of entire peoples in the interests of frontier defense was a normal part of Roman policy. That policy was evidently continued even after the departure of the Roman administration of Britain at the beginning of the fifth century. In A.D. 410 the Emperor Honorius told the cities of Roman Britain to look after their own defense; this they did by continuing to hire mercenaries. The foreign levies now came not from the distant frontiers of the empire but from the barbarian shores around the North Sea, the traditional homeland of the Anglo-Saxon peoples.

The presence of Anglo-Saxons among the Romanized Celts of Winchester shortly after A.D. 400 is indicated by the



WINCHESTER LIES beside the River Itchen near the center of Hampshire, some 12 miles to the north of Southampton and 60 miles southwest of London, the city that eclipsed it.

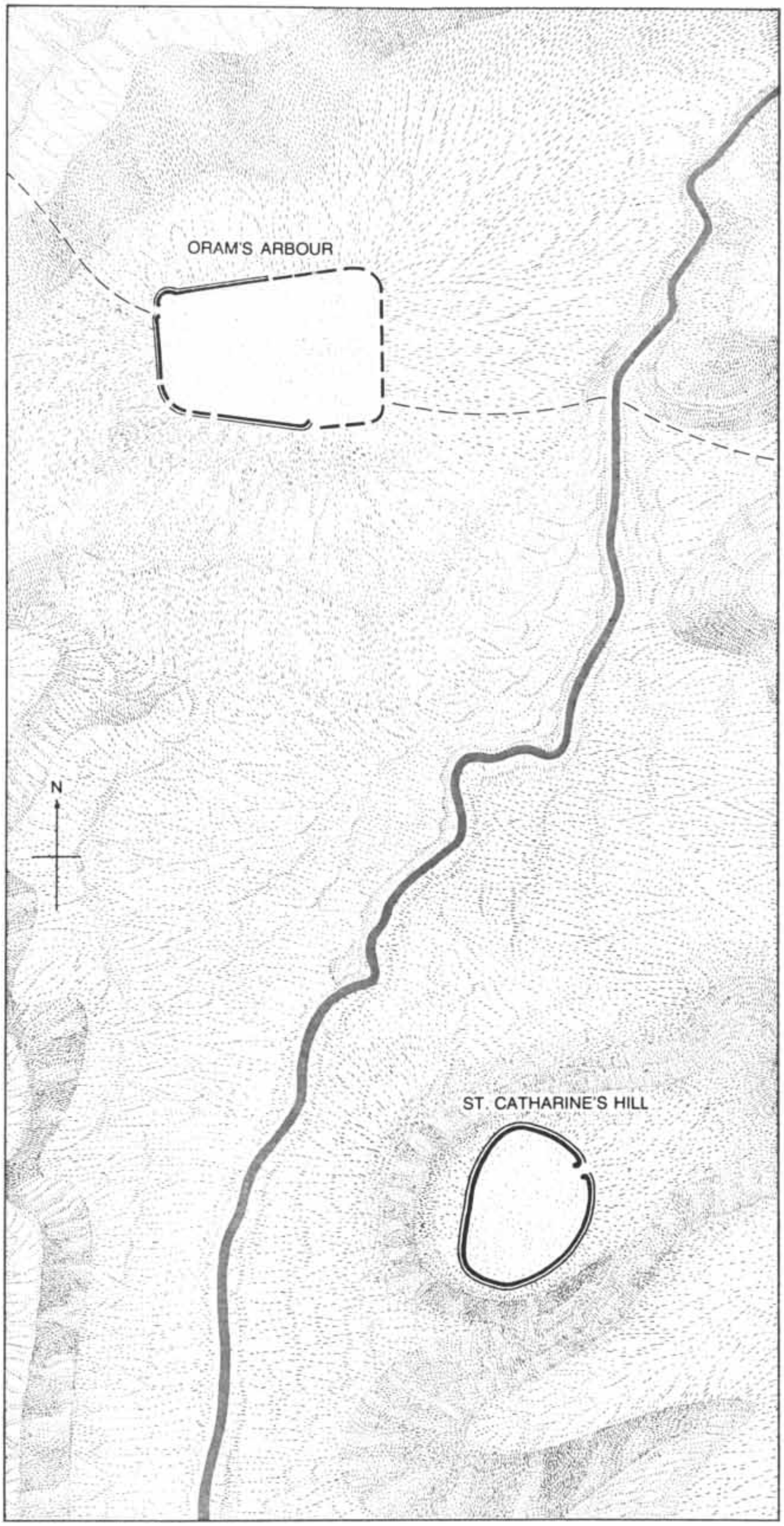
presence of pottery identical with the pottery used in their home settlements along the lower reaches of the German rivers Weser and Elbe. Together with the evidence from the cemeteries, the pottery shows that the population of late Roman Winchester was already mixed and that before the end of Roman Britain the first forerunners of the English were already established in and around the city that was eventually to emerge as the capital of an English kingdom: Wessex.

It appears that urban life in Roman Winchester came slowly to an end during the fifth century. At every site of excavation the evidence of decay, the abandonment of buildings and the loss of streets is the same. No objects and only a little pottery of the period from A.D. 450 to 650 have been found. The ruined Roman city nonetheless remained an important focal point as Anglo-Saxon settlement of the region progressed, and this is indicated by two lines of evidence.

First, comparatively few Anglo-Saxon cemeteries of the pagan period from the fifth to the seventh century have been found in the county of Hampshire. The most striking cluster of these cemeteries lies in the immediate area of Winchester, outside the city walls to the east and west. Moreover, an early cemetery that was in use by A.D. 500 lies two miles upstream from the city. Unless the former Roman center was still a focus of some kind, why should these cemeteries, each of which presumably holds the burials of a farm or a small village, be concentrated in its vicinity?

Second, in about A.D. 648 Cenwalh, king of Wessex, founded a church dedicated to St. Peter and St. Paul inside the still standing Roman walls of Winchester. About a decade later this church became the see of the bishop of Wessex. Cenwalh had built the Winchester church not as a bishop's see but simply as a minster, that is, a church served by a group of priests and clerks not necessarily living under monastic rule. One may ask what function and what community this church in the center of a ruined city was intended to serve.

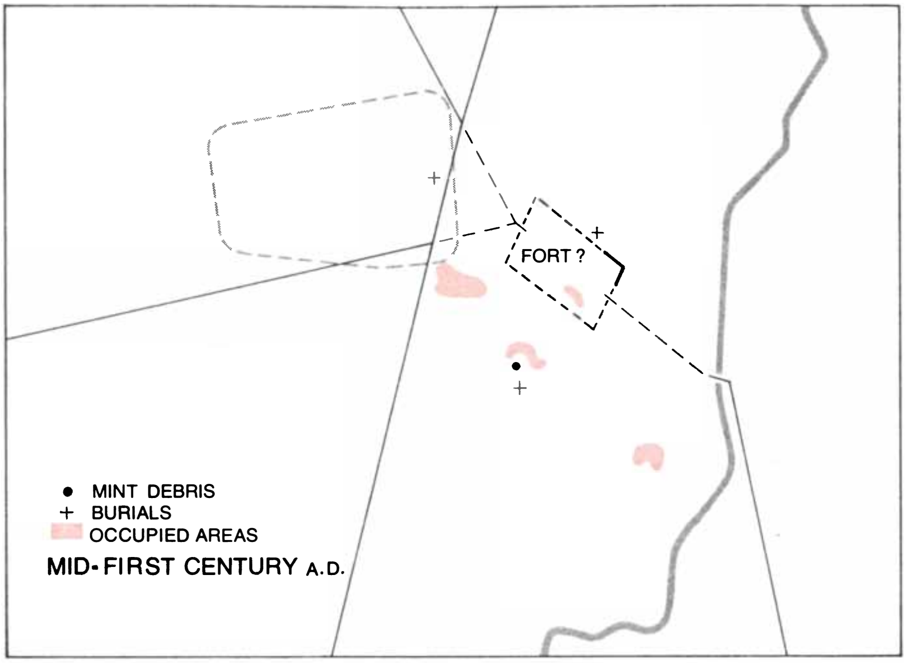
The Anglo-Saxon cemeteries suggest that from the late fifth century into the seventh Winchester was still an important place. (The Saxons transformed the name Venta into Wintancæster.) King Cenwalh's church may indicate the nature of that focus. Excavation of the seventh-century church has shown that it lies adjacent to the Roman forum. In the late Saxon period the Anglo-Saxon royal palace was immediately west of the church and was intimately associated



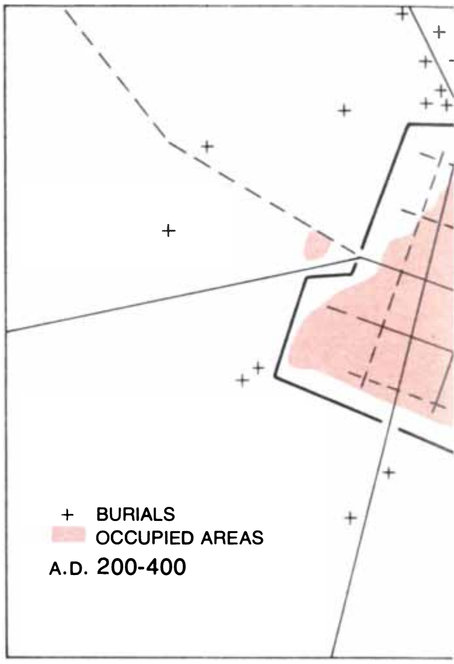
EARLIEST SETTLEMENTS in the Winchester area were two Iron Age defended enclosures. The earlier of the two (*bottom*), the hill fort now known as St. Catharine's Hill, was built in the third or second century B.C. By the time it was abandoned a second defense, twice the size of the first, had been constructed to the north on the opposite side of the river. Commanding the east-west route across the chalk downs, the new enclosure, now known as Oram's Arbour, has yielded pottery made in Italy and large bronze coins minted in Egypt. Both the route location and parts of earthworks shown at Oram's Arbour are hypothetical.

with it. The royal palace also lay close to the forum, in particular to the south end of the basilica, the principal public building of Roman Venta. The origins of the royal palace are unknown, and the site has not been excavated. The relationship of these structures may nonetheless support the following interpretation.

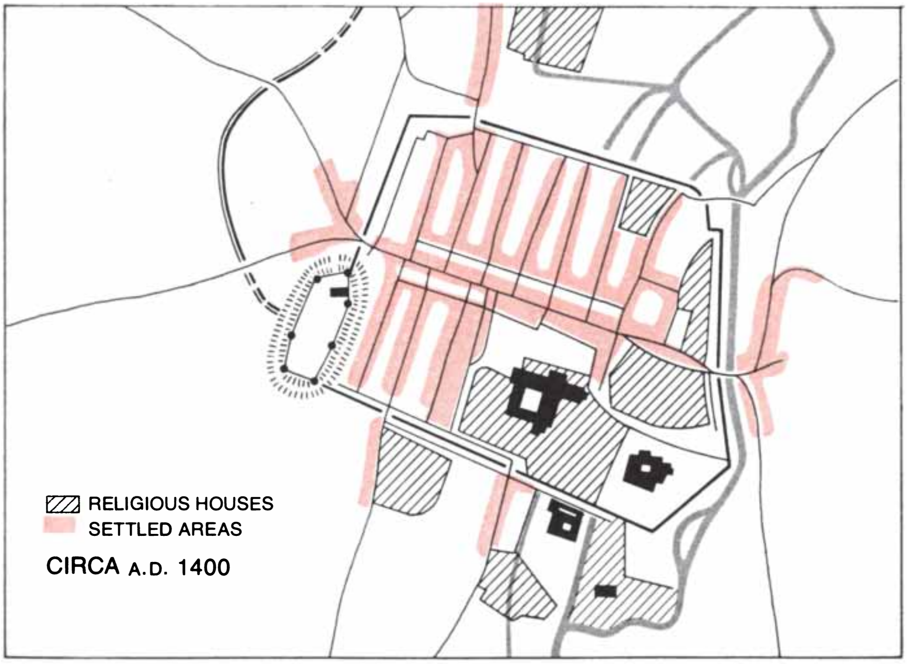
We know that the Germanic mercenaries serving in post-Roman Britain revolted against their native masters, thereby destroying the fabric of the life they had been engaged to protect. Power passed to the victorious rebels, who did not entirely forget, even when they were augmented by successive waves of settlers from barbarian Europe, that they were in some sense heirs of Rome. Continental analogies, for example the towns of Trier and Cologne, show that the buildings that had been the seat of Ro-



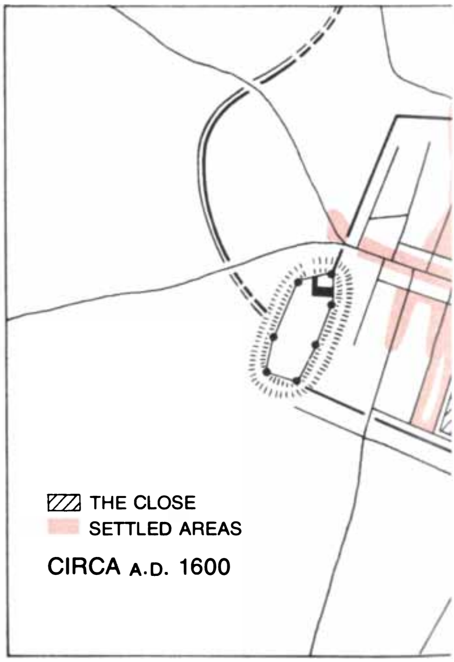
THREE MAJOR CYCLES of urban efflorescence at Winchester are illustrated on these two pages. The city's first roots were planted (*above*), perhaps in the form of a military post, to the east of the abandoned Iron Age enclosure (*light gray*) at the juncture of five Roman roads; this took place in the middle of the first century B.C. Broken lines indicate conjectural restoration.



VENTA BELGARUM, as the growing settlement was known in Roman times, was a fortified town with its streets comprising a rectangular

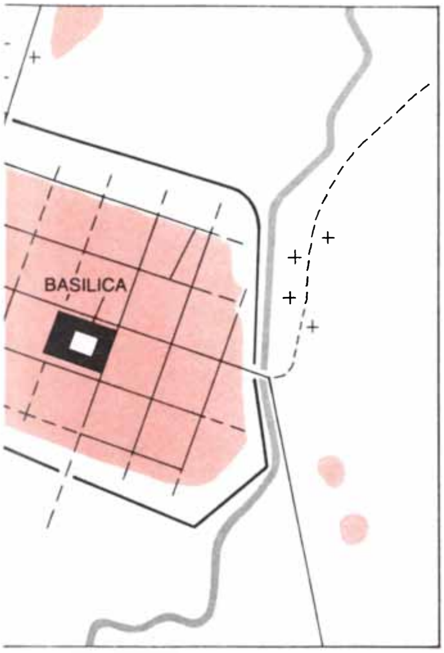


MEDIEVAL REVIVAL, first under English and then under Norman rule, raised Winchester above its former eminence as a Roman town. As seen here, some two centuries after the Conquest it had already begun to diminish in importance in spite of the great new Norman cathedral that had obliterated Old Minster. Growth of city in this period is illustrated on next two pages.

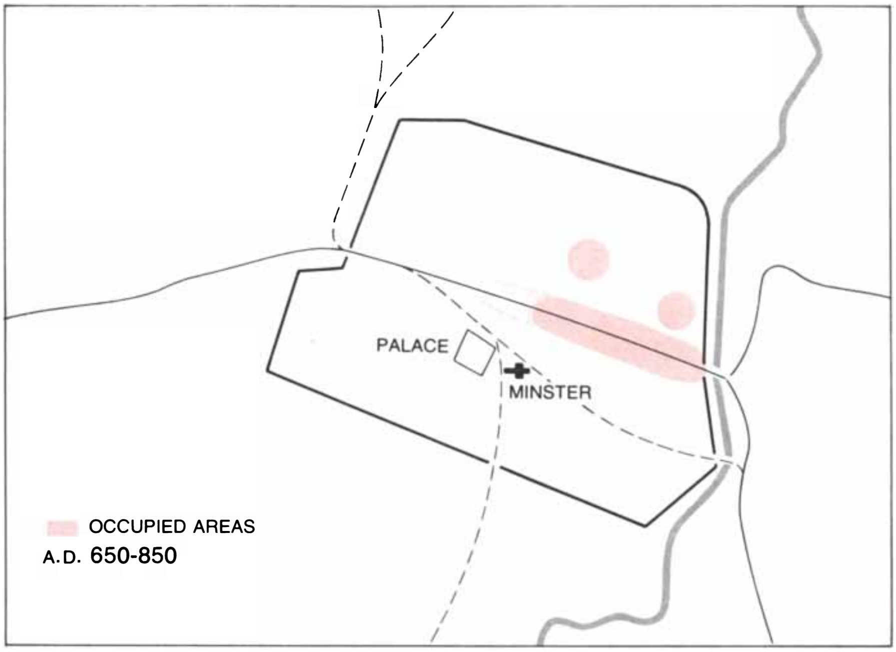


SECOND SLUMP saw the city drop to 37th place among English towns by the 1520's. Winchester fell even lower after England's monasteries man authority sometimes survived as the residences of the new rulers. The same may have happened at Winchester. Germanic peoples of Saxon origin were established in the Roman city before its collapse. To them authority over the city and its lands may have passed by conquest or by survival, and their leaders, later to be kings, may have taken up residence in or next to the basilica that was the symbol of that authority.

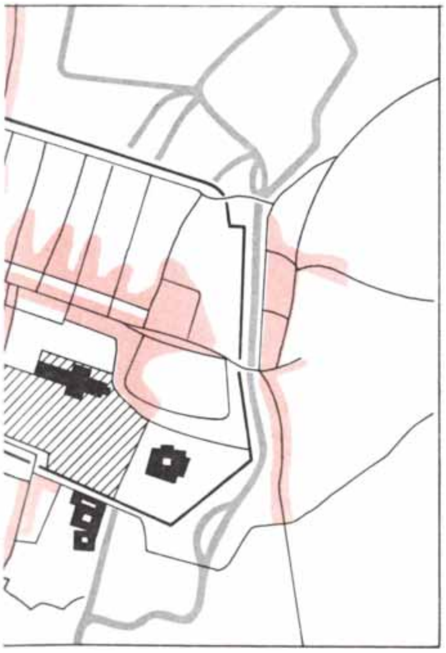
On that hypothesis the Anglo-Saxon cemeteries outside the walls would reflect the presence of this ruling element within the old walled city, and the founding of the church by a king of Wessex in the middle of the seventh century would represent the establishment of a chapel to serve the royal household. The lack of archaeological material from this period is negative evidence of a certain



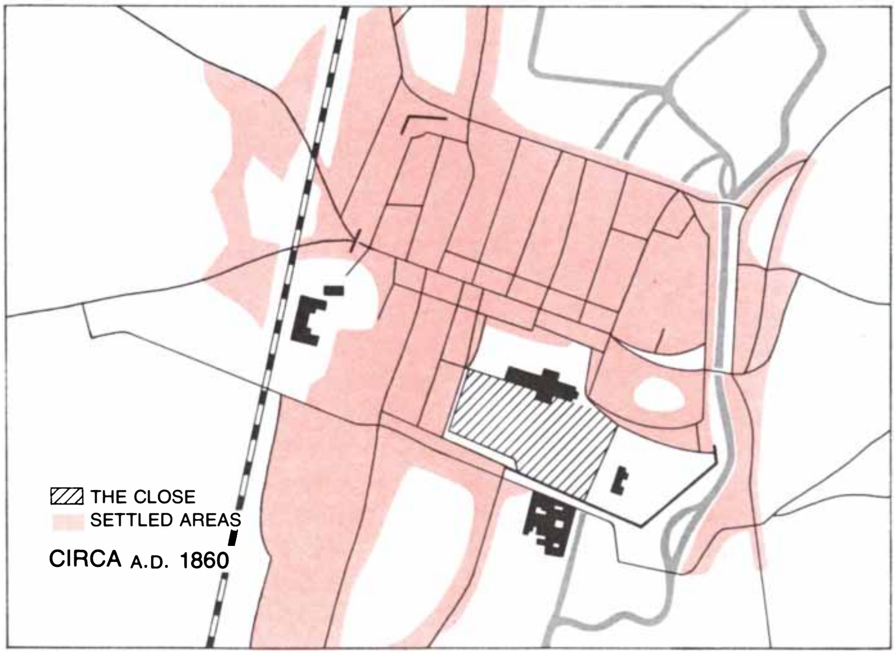
grid surrounding a central basilica and forum. In the third and fourth centuries A.D. it was the fifth-largest of the settlements in Roman Britain.



FIRST DECLINE came in the centuries after the end of Roman rule. By the sixth century only the wall and a single road bisecting the enclosed area remained of Venta Belgarum. For the next three centuries the town, called Wintancæster by the Saxons, may have sheltered a "king's hall" near the ruined basilica; the bishop's minster nearby could also have served as a royal chapel.



were dissolved during the 1530's. Depressed for 300 years thereafter, the cathedral city largely marked time until early in the 19th century.

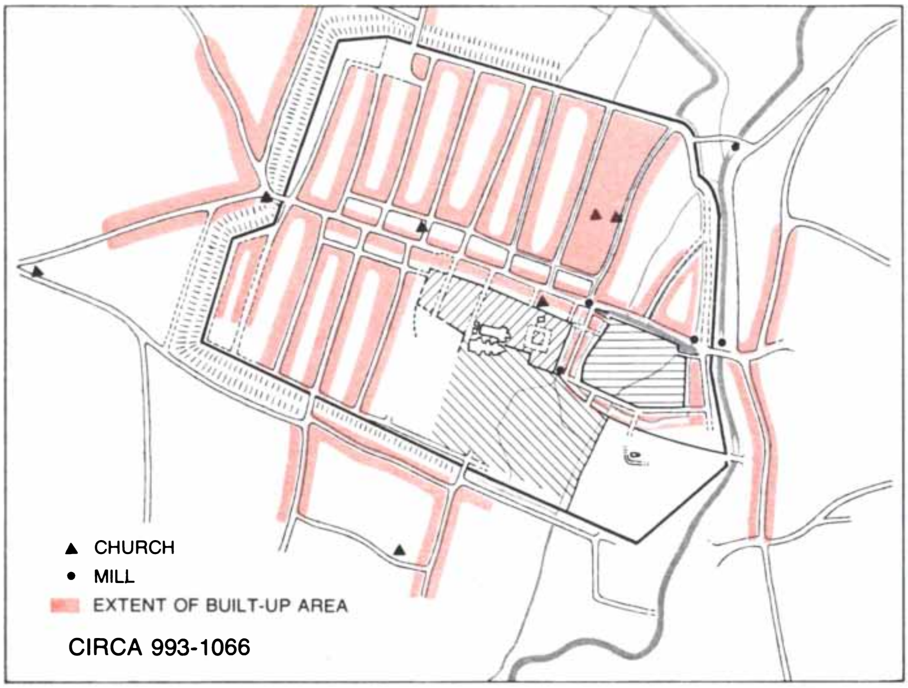


VICTORIAN REVIVAL of Winchester included construction of a railway just west of the city, which now stretched beyond the bounds of its Roman walls although only slightly exceeding its maximum extent in medieval times. The southeast quadrant of the old walled area remained dominated by the cathedral and its "close." Just south stands Winchester College, a public school.

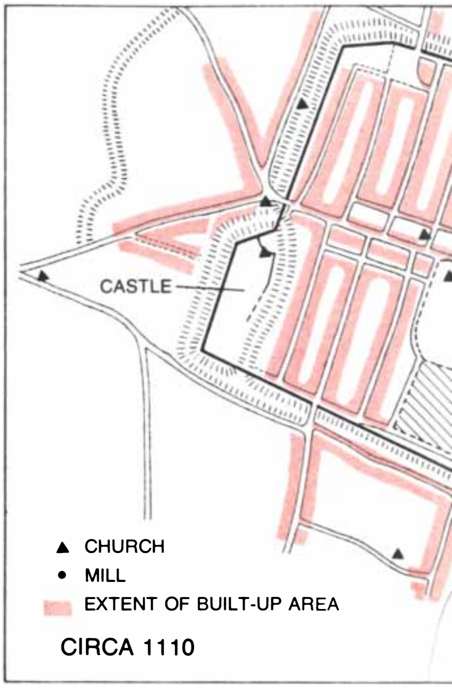
value: it shows that the greater part of the walled area was uninhabited. On the other hand, it has nothing to say about the Anglo-Saxon royal palace that has been buried under the cathedral graveyard for eight centuries and remains unexcavated.

With the founding of the church, which later became known as Old Minster, the city of Winchester entered a new phase. A few contemporary written records, an increasing amount of archaeological evidence and comparisons with other English and continental centers make it possible to present a less hypothetical picture of the character of the city in the two centuries following A.D. 648.

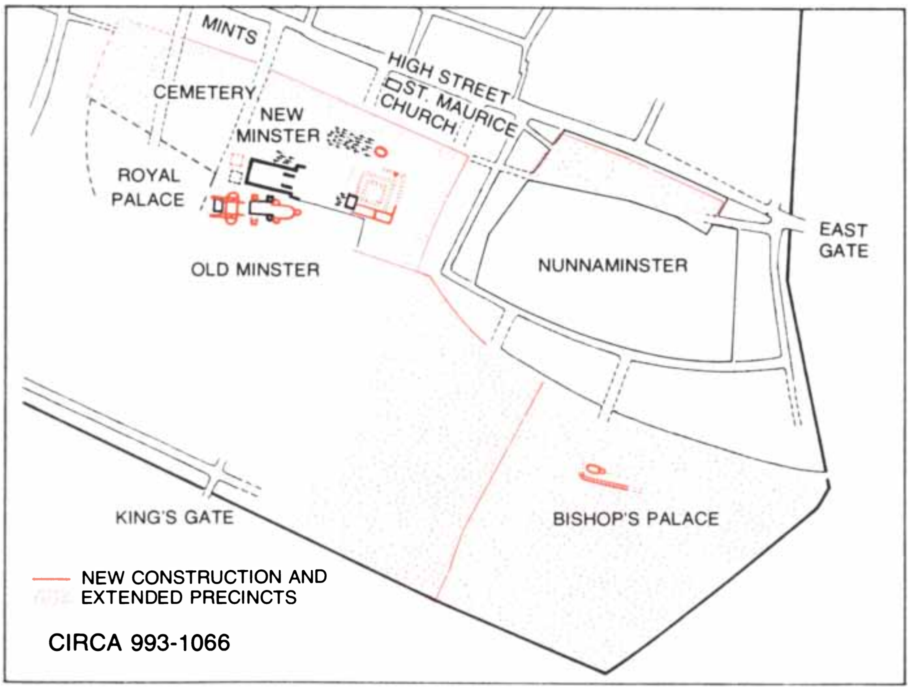
Within the walled area of Winchester four components become evident. One is the bishop's church and its community.



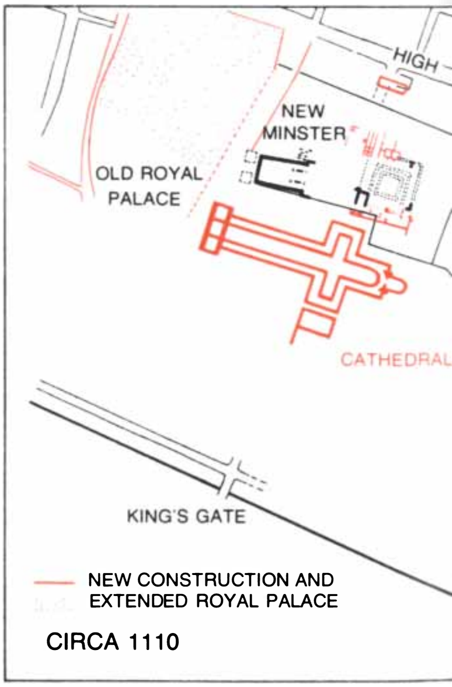
ANGLO-SAXON WINCHESTER just before the Conquest was a town with regularly aligned streets, elaborate defenses and many churches in addition to the Old and New Minsters and Nunnaminster. The illustration directly below shows the city's southeast quadrant in detail.



POST-CONQUEST DECADES were notable for Norman expansion. A new castle (*lower left*) enhanced the



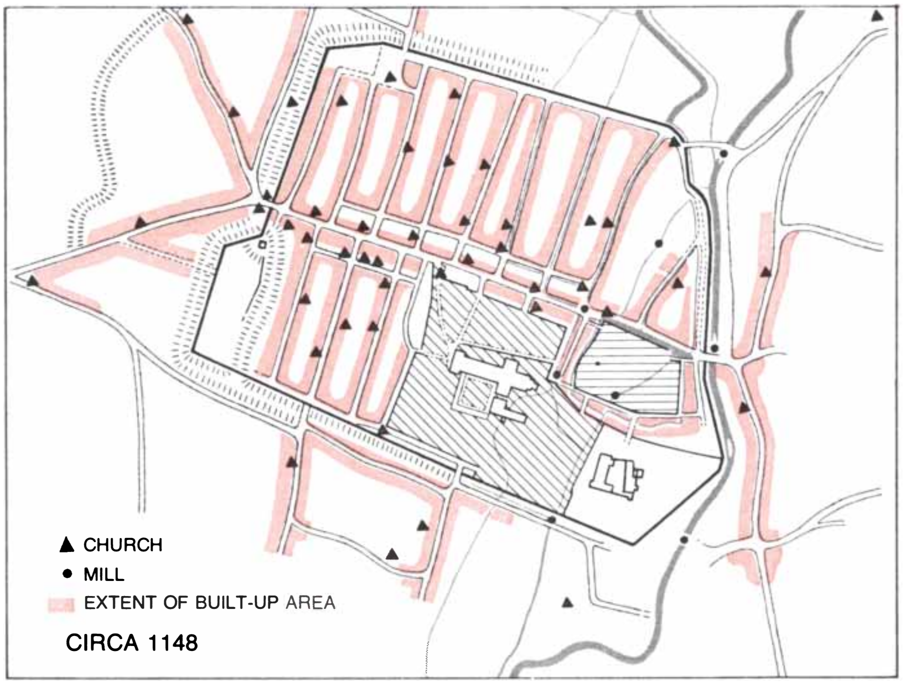
SOUTHEAST QUADRANT of Winchester at the close of the Anglo-Saxon era was the site of the royal palace, the bishop's palace and, in addition to lesser churches, the cathedral church or Old Minster, New Minster and Nunnaminster (nuns' church). Mints may have been located here.



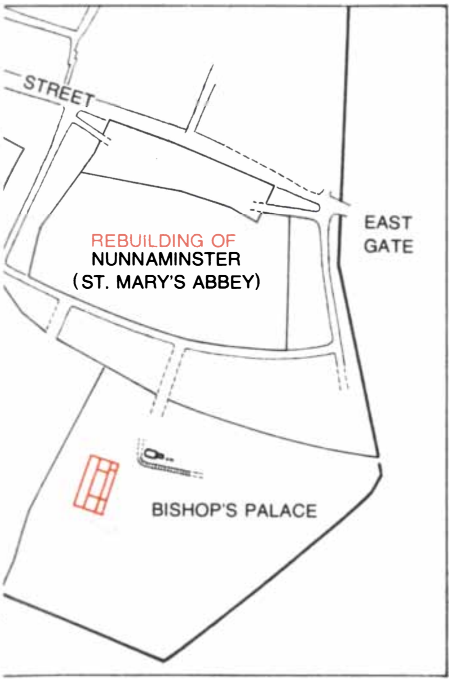
NORMAN CHANGES doubled the size of the royal palace, rebuilt and extended the bishop's palace and also

Another is the royal residence; there is more circumstantial evidence for its existence during this period, when the church was the burial place of the Wessex kings. A third component is the presence of an unknown number of private residences; there is evidence of two such residential complexes. Of one, only the name survives as a description of an area within the city's East Gate: Coitburi. Names of this type, the second element signifying a defensible enclosure, are known from early London. The other private residence has actually been excavated in part; its earliest feature is a small private cemetery of the seventh century, probably adjacent to the earliest buildings, which remain unexcavated. The area of the cemetery was eventually built up; the first structures here were of timber, and in about A.D. 800 a stone

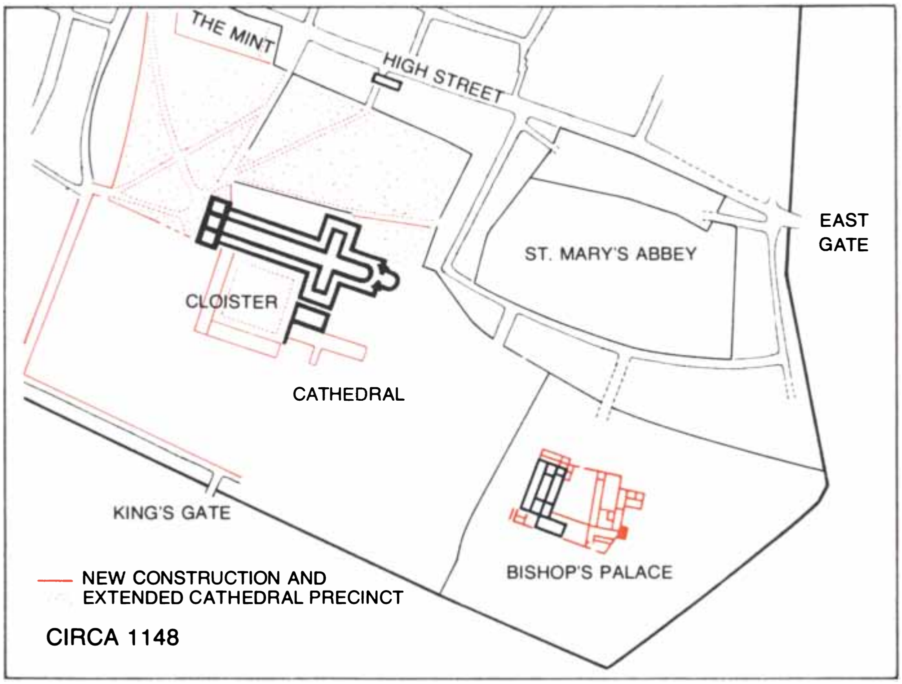
defenses and a great new cathedral rose. As a Norman seat the city stood second among English towns.



NORMAN APOGEE at Winchester came in the 12th century, when the city's churches numbered more than 50. Winchester's decline began that same century with a loss of close contact with the court. The trend was accelerated by removal of the royal treasury to London during the 1180's.



raised a great new cathedral. At the time of completion the cathedral was the longest church in all England.



CATHEDRAL PRECINCT had extended over the sites of the royal palace and New Minster by 1148. Nunnaminster, now called St. Mary's Abbey, was rebuilt once again. The bishop's palace, one of the greatest houses of its age, had by now almost reached its ultimate dimensions.

house was built [*see illustration on opposite page*]. Its remains contain evidence for the working and assaying of gold. The wealth of the burials in the private cemetery (one contained a necklace with gold and garnet pendants and 27 silver rings), the construction of a stone building and the working of precious metal all suggest occupants of high social status.

The fourth component is more problematical. It seems likely that the royal, ecclesiastical and private residences must have been supported by some service population, perhaps in the form of a developing street market along the eastern part of High Street, in the area known by about A.D. 900 as *ceapstræt,* or market street. As extensive excavations have shown, much of the walled area of the city was certainly uninhabited at this time, but contemporary records and the archaeological evidence both indicate that some of the walled acres were used for the grazing of livestock and the raising of crops.

The existence of these four components does not make Winchester an urban community at this time; there is no evidence of industry, of a dense population, of trade or of a full social hierarchy. Indeed, only the higher levels of society seem to be present. Comparisons with Anglo-Saxon Hamwih (modern Southampton), some 12 miles downstream, are instructive. At Winchester we have a royal residence and royal burials, a bishop and his church, the homes (two at least) of subjects of substance and just possibly a mint. Southampton had a mint but no other obvious marks of social greatness. At this time, however, Southampton had a substantial population, much industrial activity and the elements of a regular street plan. Moreover, there is evidence of long-range trade. In contrast, none of these is found in Winchester.

Here are two different kinds of settlement: the old royal and ceremonial center, limited in its extent and functions, and the complementary port and industrial settlement emerging as a true town at the head of a superb natural harbor. Now, Anglo-Saxon kings had many residences, and the primitive apparatus of government moved with the king as he journeyed among his estates. Winchester therefore was not at this time a capital any more than it was an urban community. It was, however, one of the more important royal residences, perhaps because of its church and the city's close association with the royal house.

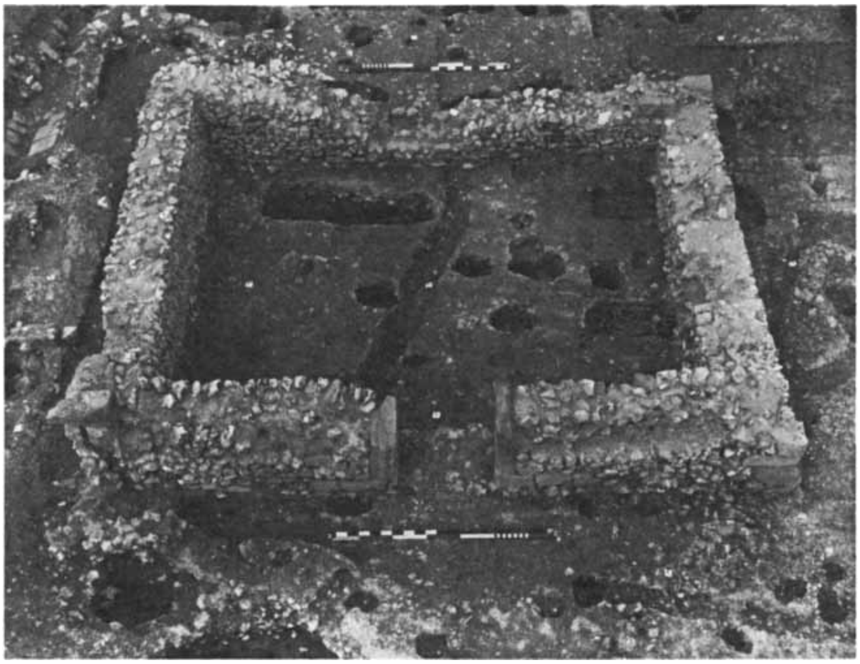
Beginning toward the close of the eighth century England was subjected to a series of Viking attacks that steadily increased in severity. By the middle of the ninth century these had evolved into a phase of extensive Scandinavian settlement. Alone among the English kingdoms Wessex survived. After the Battle of Edington in 878 and the Treaty of Wedmore in the same year King Alfred, who ruled Wessex from 871 to 899, set about bolstering the defenses of his kingdom. Alfred's strategy was based on a series of "burhs," or fortified places, so located that no part of Wessex was more than 20 miles from one of them. The burhs were of several kinds: simple forts, refurbished Roman fortresses, newly created towns and former Roman towns that had been refortified and replanned. Winchester was in the last category.

The Roman defenses of the city were brought back into commission and the city gates repaired or rebuilt. Within the walls a new street system was laid out along the axis of the main east-west street; this street had survived in modified form from the Roman period. The other new streets, however, had no connection with the Roman pattern, which had long since vanished. The elements of the new pattern—the east-west High Street, the back streets parallel to it and the intersecting north-south streets—

ROW OF SMALL DWELLINGS, built in the 13th century, lies exposed by excavation. Only the earth floors and the clay sills of walls have survived. Medieval records place cloth-finishing works in this area; the cottages may have housed workers in the industry.

have remained in use with minor changes down to the present. A fourth element, a street running around the entire city inside the wall and providing direct access to the city's perimeter defenses in time of war, is partly lost today. As its original function became unnecessary it was built on in many places.

The ninth-century street plan shows that the entire walled area of the Roman city—143 acres—was brought back into use by the end of the century. Similar street systems can be seen in many of the other burhs set up by Alfred, and there can be no doubt that they represent a deliberate intent to establish urban communities; this English episode of organized town foundation is without parallel in early medieval Europe. The blocks formed by the new streets seem to have represented land apportioned for permanent settlement. In such places military effectiveness was to be secured by economic success.



ANGLO-SAXON HOUSE, some 23 feet square, may have been built around A.D. 800. The four corners and the doorway incorporate masonry, much of it taken from Roman buildings, but the walls are mainly courses of flint rubble. The two graves (*right*) continue under wall; antedating the house construction, they form part of a cemetery of seventh century.

Not all Alfred's burhs were successful, but Winchester never looked back. By the 960's the privacy of its monasteries had to be protected against a rising tide of urban life. Before the end of the century several city streets were named after the trades practiced in them. There was a Tanner Street, a Fleshmonger Street, a Shieldmaker Street and later a Shoemaker Street (to give their names in modern English), and suburbs were growing outside each of the city's five gates.

The southeastern quarter of Winchester gave the city its unique character. Here, 100 years before the Norman Conquest, was the most remarkable group of royal and ecclesiastical buildings in Anglo-Saxon England. Edward the Elder (899–924) founded New Minster and Nunnaminster, that is, a "nuns' minster." In the reign of King Edgar (959–975), Old Minster, New Minster and Nunnaminster were reformed and reconstructed. The bishop's palace was established in the same quarter of the city at about the same time, and between 971 and 994 Old Minster was entirely rebuilt. By this time written evidence at last confirms the existence of the royal palace immediately east of the cathedral.

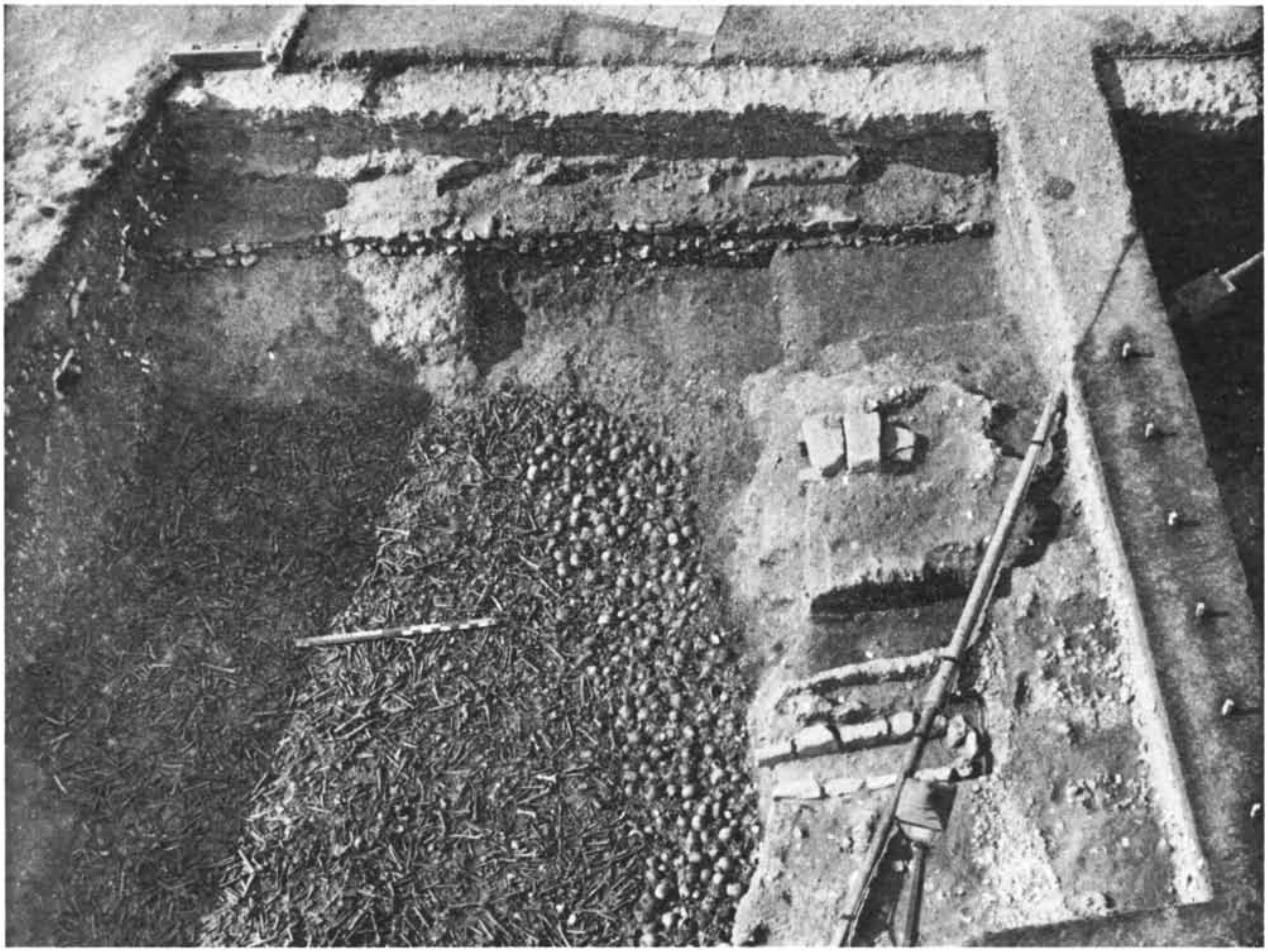
It was in these buildings toward the end of the Anglo-Saxon period that the apparatus of a centralized English state began to emerge. By the reign of Cnut, or Canute (king of both England and Denmark from 1016 to 1035), Winchester had become the permanent repository of the king's treasure. The time of Edward the Confessor (1042–1066) may have seen the emergence of an embryo financial and secretarial administration. The cathedral continued its ancient association with the ruling house: Cnut was buried in it in 1035 and Edward was crowned there at Easter, 1043, thus formalizing his accession the preceding year. Even before the Norman Conquest the custom seems to have been established of the king of England's wearing his crown in Winchester Cathedral at Easter, the most important feast of the Christian year.

No other place in England and few places in all Europe played such a central role in the life of a state during the 11th century. Yet Winchester was not the largest city in the realm. It was perhaps fourth in size and economic power, being surpassed by London, York and Lincoln in that order. Yet it was emerging as a kind of national capital, a distinction that was destined to pass to Westminster and London in the century that followed.

Victorious at Hastings on October 14, 1066, William the Conqueror seized Winchester without opposition in November, opening the way for the surrender of London and his coronation in Westminster Abbey on Christmas Day. The effects of the Conquest on Winchester were as complex as they were considerable. In the larger buildings, in the composition of the upper levels of urban society and in social fashion there were profound changes. In administration, in the bulk of the population and in the basic fabric of the houses and the streetscape there was essential continuity. If for many in Winchester the immediate dislocation caused by the Conquest and the appropriation of land for new buildings was serious, the massive Norman financial investment in major public works during the remainder of the century and the presence of royal officials, barons and magnates of the newly rich Anglo-Norman aristocracy ensured for the city a rapid recovery and a clear improvement in its wealth and status.

By about 1100 the ancient role of Winchester as a royal center had been given new emphasis. In February, 1067, a Norman castle had been begun at the point where the Roman defenses at the southwest corner of the city formed a salient. In about 1070 the Anglo-Saxon royal palace was extended northward to High Street and doubled in area, the additional space being required for the construction of the Conqueror's hall and palace. East of the palace the total rebuilding of Old Minster was begun in 1079. The eastern part of the new cathedral was dedicated 14 years later, in 1093, and the entire project was completed in 30 or 40 years.

The Norman cathedral demonstrated most clearly, as perhaps its builders had intended, not only the eminent role of the city but also the finality of the Norman acquisition of the Anglo-Saxon



MORE THAN 1,100 REINTERRED SKELETONS were found near the west end of the cathedral. In digging foundations for the Norman cathedral the builders disturbed many burials. In filling the trench dug to gather stone from Old Minster for use in the new building they disposed of the disturbed remains. Skulls were placed toward west according to tradition, but other bones were jumbled.

state. The new cathedral was more than 500 feet long, making it larger than any other church in England or Normandy, longer than old St. Peter's or any of the churches on the pilgrims' route to Santiago de Compostela in Spain. Only the contemporary abbey church built by St. Hugh at Cluny in Burgundy was longer.

With the building of the castle, the rebuilding of the palace and the cathedral, the repair and reconstruction of New Minster and Nunnaminster (renamed St. Mary's Abbey), Winchester in about 1100 was a principal residence of the Norman kings, the seat of the royal administration and a center of great ecclesiastical importance. Englishmen had yielded place to Normans in the houses along the most important streets of the city. The English had also adopted Norman fashions to such an extent that 70 percent of the citizens' names recorded in about 1110 were foreign, whereas only 15 percent had been before the Conquest. The English were nonetheless still prominent in affairs. Winchester's mint was now second in importance only to London's, and the moneyers, whose ranks included the leading burgesses and property owners in the city, were almost all English.

Winchester probably reached its zenith in the early years of the 12th century. After 1104 Henry I abandoned the custom of the annual Easter crown-wearing at the cathedral, a practice that had been regularly observed by his predecessors. Royal interest shifted from the palace in the center of the city, in intimate contact with the cathedral, to the new castle on the hill beside the wall. By the 1130's the palace was no longer a royal residence; it may by then have passed to the bishop, whose role in city affairs was now increasing. At that time the rebuilding of the episcopal palace at Wolvesey in the southeastern corner of the city was undertaken. Successive kings also extended the period over which the bishops of Winchester might enjoy the profits of St. Giles Fair, held on the hill east of the city. The three days originally granted by William Rufus in 1098 were increased to 16 days under Henry II. Although the fair was probably of pre-Conquest origin, its heyday came in the 13th century, when it was one of the most important fairs in England and was attended by traders from many parts of Europe.

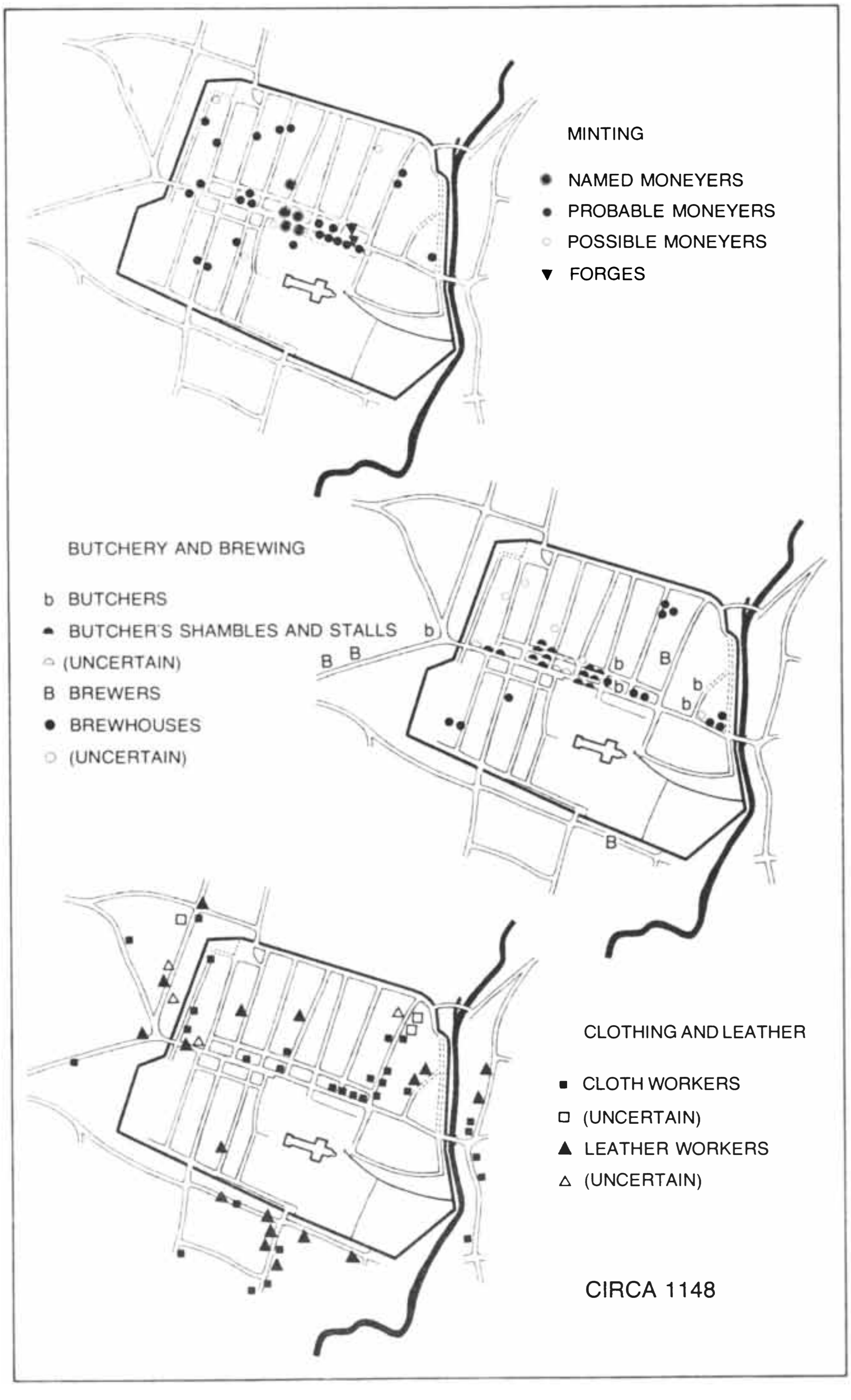
In the civil war of 1141 Winchester was seriously damaged. The old royal palace was burned down, and St. Mary's and Hyde Abbey and many parish churches and private houses suffered severely when the city was sacked by the London contingent supporting the king. By that time London had been the largest and wealthiest city in England for some 200 years. Westminster had emerged as a royal residence in the 11th century and increased greatly in importance with the rebuilding of the Abbey by Edward the Confessor and his burial there in 1066. By the middle of the 12th century an increasing number of administrative functions were located at Westminster. Finally in the 1180's even the tradition of Winchester as the site of the royal treasury gave way, and the king's treasure was transferred to London.

The close link between Winchester and the crown was now severed. The castle remained an important royal residence, often embellished and often visited, but it was of no more importance than many another great house. The economy of the city held up during the rest of the 12th century, but there are signs of trouble in the 13th century, as first the western suburb and then the western neighborhood within the walls began to decline. Large areas of the city passed into religious hands. By the 14th century considerable tracts within the walls were no longer built up. A petition of 1440 cited the destruction of 11 streets, 17 parish churches and 987 houses as a result of pestilence and the withdrawal of trade. Where Winchester had occupied second place among English cities at the end of the 11th century, by 1200 it was sixth or lower. By 1334 it was 14th, by 1377 it was 29th and by 1527 it was 37th. Many of the city parishes were amalgamated in the early 16th century, but it was the suppression of the monasteries in 1536–1539 that wrought the greatest changes, removing three monastic communities, four friaries and several lesser institutions.

The built-up area of the city was by now confined to the central and eastern parts of High Street, to the adjacent areas of the side streets, to the main north-south street and to the eastern and southern suburbs. So it was to remain for three centuries. By the early 19th century a revival had begun, encouraged by the growing role of Winchester as a garrison town and by the advent of the railway in 1839.

Ancient Winchester now lies under the streets and buildings of an active and dynamic modern urban center. Reconstruction, redevelopment and the redesign of approach roads and internal streets are destroying the evidence of the city's past at a quantifiable rate. The pattern of the city's Roman-built defenses was effectively breached for the first time only in 1939. By 1950, 2 percent of the defenses had been destroyed and by 1965, 8 percent. By 1980 completion of the city's traffic plan will have raised this figure to 35 percent. A third of the 2,000-year-old defensive system will have been removed in 40 years. There are many similar examples.

In such a situation the raw material for the study of urban evolution has to be rescued now or not at all. Winchester is exceptionally rich in written records, but they barely touch the first 1,000 years of the community's existence, its Iron Age and Roman cycles and its Anglo-Saxon rebirth. Historical data only become full during the time of Winchester's long medieval decline. This is a pattern that is repeated all over Europe. The basic evidence for the study of urban origins and growth, for the waxing and waning of our towns and cities, has not been recognized until the last moment before its destruction. In London not more than 15 years remain in which to undertake an inquiry that will never again be possible. The example of Winchester may show what can be won. It also shows how much may be lost.



URBAN TRADES AND INDUSTRIES in mid-12th-century Winchester included victualing and manufacture in addition to the minting of coins. Five properties, four of them on High Street, were occupied by moneyers whose names appear in an 1148 survey. Another 27 properties are identifiable as probably or possibly moneyers', and two moneyers' forges are known.

Visual Pathways in Albinos

A genetic abnormality in Siamese cats, white tigers and other albino mammals provides a natural experiment for investigating how the brain acquires an orderly picture of the outside world

by R. W. Guillery

Siamese cats, white tigers, pearl mink and albino rats all exhibit a family of genetic anomalies in which some form of reduced pigmentation is combined with a congenital abnormality of the central visual pathways: some of the optic nerve fibers go to the wrong side of the brain. This abnormality, which can be associated with crossed eyes, has provided new ways of studying the visual pathways, which in recent years have been investigated from two very different points of view. One approach has been to determine how the brain analyzes the visual signals it receives from the eyes [see "The Visual Cortex of the Brain," by David H. Hubel; SCIENTIFIC AMERICAN, November, 1963]. The other has been to discover the developmental processes that make the orderly connections characteristic of the visual pathways [see "The Origins of Nerve-Cell Specificity," by Marcus Jacobson and R. Kevin Hunt; SCIENTIFIC AMERICAN, February, 1973].

The abnormality of the visual pathways has been most thoroughly studied in Siamese cats. I shall first describe what happens to the visual pathways in these cats and how the brain deals with the abnormal visual input. Then I shall discuss some of the broader developmental problems that arise when one considers how a single identifiable gene might produce an abnormal set of connections between the retina and the brain.

Some years ago, during a study at the University of Wisconsin Medical School of the central visual pathways in cats, I found that in one brain the lateral geniculate nucleus, the main cerebral cell group responsible for relaying messages from the retina to the cerebral cortex, was strikingly abnormal. In a normal lateral geniculate nucleus there are clearly differentiated layers of nerve cells. In the abnormal brain the stained cells formed layers that were disrupted and to some extent fused [*see illustration on page 48*]. Such an abnormality was of particular interest because the experimental animal in question was a Siamese cat and Siamese cats are commonly cross-eyed. It seemed possible that the odd structure of the geniculate layers might be related to the abnormal alignment of the eyes, since it is in the lateral geniculate nucleus that the inputs from the two eyes are matched and passed on to the cortex, and since correct alignment of the eyes is likely to require a properly matched input to the cortex.

In a normal brain this matching of inputs is accomplished by the layered organization of the two geniculate nuclei, each of which receives an orderly input from one half of the visual field [*see illustration on page 49*]. The left hemifield is represented in the right nucleus, which in turn projects to the right visual cortex; the right hemifield is represented in the left nucleus (and cortex). Each nucleus is supplied by fibers from ganglion cells in both retinas, so that the parts of each visual hemifield that can be seen with both eyes are represented at least twice, the input from one eye in the first layer called Layer *A*, and the input from the other eye in the next layer, A_1. This would be the simplest case, in which there are only two layers; actually most mammals have more than two, and these several layers always form two sets, each set receiving an orderly pattern of inputs from one eye. Each layer thus receives a map of one visual hemifield, and the maps in the several layers are always in register with each other, so that it is possible to draw "lines of projection" through the geniculate nucleus, roughly perpendicular to the layers, that correspond to single points in the visual field.

The brain in which I first observed the abnormally layered lateral geniculate nucleus came from an animal in which the fibers of one optic nerve (the bundle of fibers issuing from one eye) had been cut. The cut fibers degenerated and later accepted a special stain, thus revealing which parts of the lateral geniculate nucleus had been innervated by the cut optic nerve. In a normal cat some of the degenerating fibers would have crossed over to innervate Layer *A* on the side of the brain opposite to the cut, and others would have remained on the same side but would have gone to Layer A_1 [*see illustration on page 50*]. In the Siamese cat, in contrast, some fibers that crossed over went on to innervate parts of Layer A_1, which would normally have received an uncrossed input. Conversely, on the side of the brain where the optic nerve had been cut only parts of the layers that should have received an uncrossed input (and should therefore have shown the stained degenerating fibers) were in fact innervated from the cut side. Further investigations in our laboratory and others established that all Siamese cats have this abnormality, even those that are not obviously cross-eyed. The cross-eyed cats, however, do seem to have more of the misdirected fibers than straight-eyed cats.

The pattern of fiber degeneration told us that some nerve fibers that would normally stay on their own side of the brain instead cross over to the opposite geniculate nucleus, but it did not show how the visual fields are represented within the abnormal nucleus. In order to understand the pattern of the inputs the brain is receiving from the abnormal pathways it is necessary to determine whether the fibers are merely misrouted with regard to their crossing at the optic chiasm or are more extensively misdirected, so that the orderly point-to-point representation

CROSS-EYED WHITE TIGER is an example of an albino mammal that has visual abnormalities as well as reduced pigmentation. The tiger, a female named Rewati, is normally a resident of the National Zoological Park in Washington but is temporarily at the Brookfield Zoo in Chicago. White tigers have white stripes rather than the normal orange ones alternating with their normally colored brownish-gray stripes. The eyes are blue. The unusual coloring, and the less common squint, are related to an abnormality of the brain's visual pathways that was discovered by the author's group in Siamese cats and was also found in Rewati's brother Moni.

of the retina within the lateral geniculate nucleus is scrambled.

Jon H. Kaas and I studied this problem with two methods. One was to make fairly small lesions in the retina of Siamese cats, thus destroying a patch of retinal ganglion cells so that the corresponding fibers running from the retinal cells to the lateral geniculate nucleus would degenerate. We were then able to stain the degenerating nerve fibers differentially and so determine which part of the geniculate nucleus was innervated from the damaged part of the retina. These experiments were rather crude because it is difficult to make very small retinal lesions, but they did suggest to us that the point-to-point organization of the projection from the retina to the geniculate nucleus was normal. Lesions in the upper part of the retina produced degeneration in the front parts of the nucleus and lesions in the lower retina produced degeneration in the back of the nucleus. This is the normal arrangement in cats. Apparently even the abnormal nerve fibers went to the proper parts of the lateral geniculate nucleus, albeit on the wrong side of the brain. Moreover, when the lesions were near the center of the retina, the degeneration was near the midline in the geniculate nucleus, whereas peripheral lesions produced lateral zones of degeneration. Again, this is the normal pattern. We found this normal sequence in all parts of the lateral geniculate nucleus, and we concluded that even the fibers going to the wrong hemisphere are properly specified with respect to the geniculate locus of their termination.

Electrophysiological methods provided a more elegant method of studying the routing of retinal fibers. A micro-

SIAMESE CAT has a temperature-sensitive variant of the gene that causes albinism. As a result the animal is pale in color except for the extremities, which are cooler and therefore dark. The abnormality of the visual pathways has been found in all Siamese cats examined so far. The cats' and tigers' blue eyes are unusual; fully albino animals lack eye pigment completely, and their eyes appear pink.

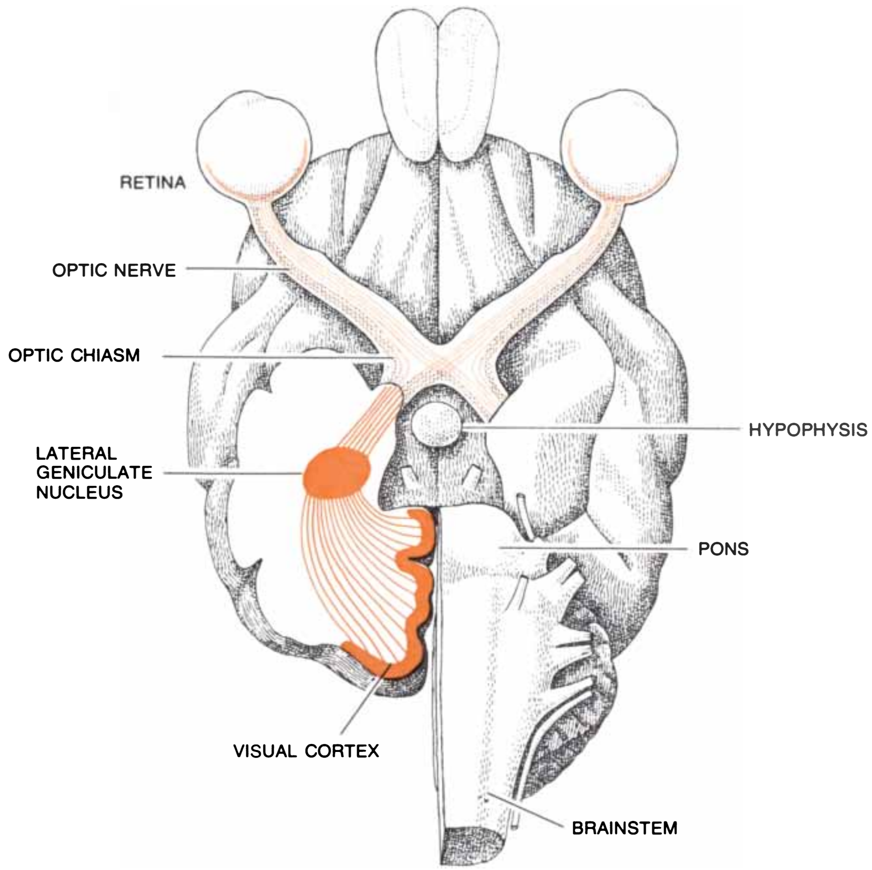
electrode lowered into the lateral geniculate nucleus of an anesthetized cat records the activity of single geniculate cells or of small groups of cells as a light is flashed in various parts of the visual field, and thus indicates what parts of the field must be stimulated in order to activate each group of nerve cells. By exploring the lateral geniculate nucleus systematically one can plot the representation of the visual field within the nucleus [*see illustrations on page 51*]. The experiments confirmed our degeneration studies. Although some of the fibers from the retina to the lateral geniculate nucleus go to the wrong side, within each nucleus the sequence of the retinal representation in Siamese cats is entirely normal. Since retinal ganglion cells lying in well-defined patches of retina send their fibers to the wrong side of the brain, however, the side-to-side order of the geniculate representation of the visual field is reversed in the part of the nucleus with an abnormal input.

In the left geniculate nucleus the visual-field sequence is mapped normally in both layers for segment No. 9 and again for No. 13 through No. 16 [*see illustration on page 52*]. In between, however, segments Nos. 10, 11 and 12 are mapped normally only in the upper layer (A); the lower layer (A_1) maps segments Nos. 5, 6 and 7 in the reverse order. The same kind of anomaly affects the right nucleus: five of the segments are mapped normally, but three segments are inserted incorrectly into Layer A_1. We can reasonably ask what this means in terms of the cat's visual abilities. We cannot, unfortunately, describe what the cat sees. In order to understand how the cat's brain deals with the abnormal input we need to know how the lateral geniculate nucleus normally passes information on to the cerebral cortex.

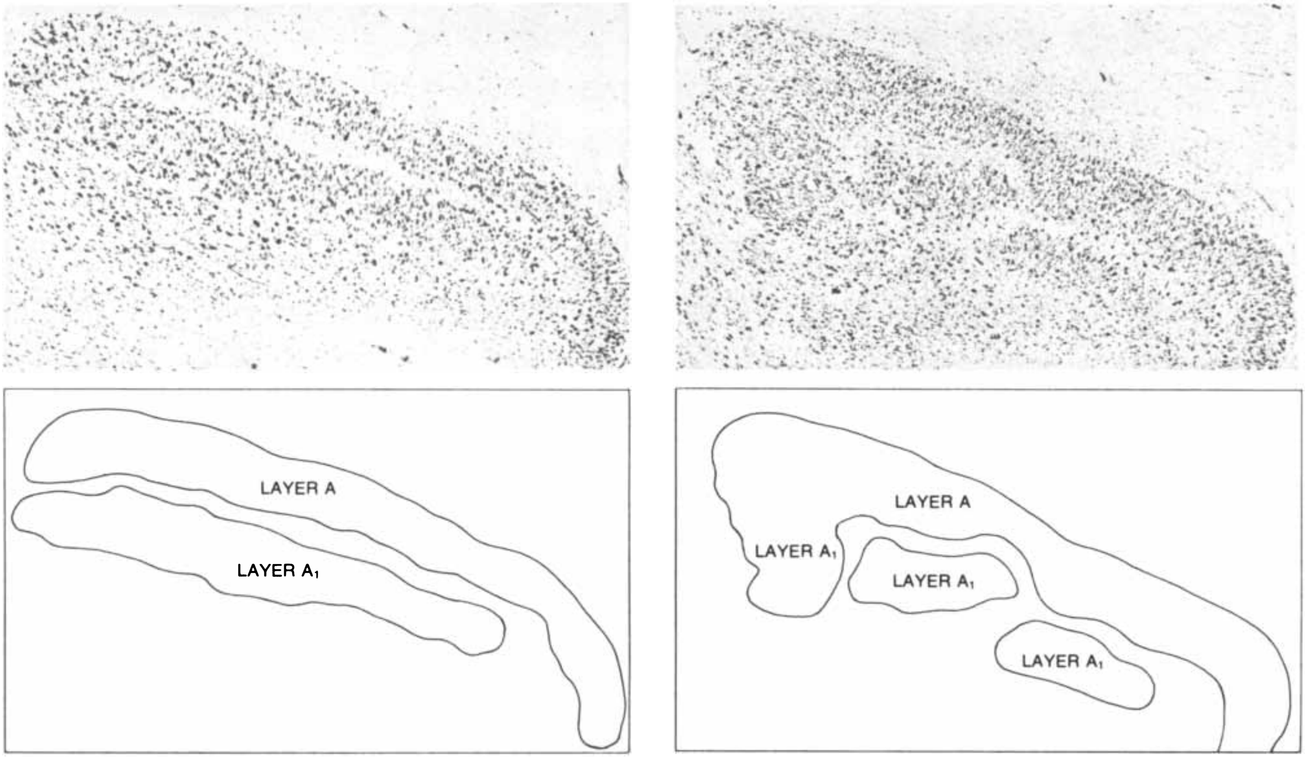
In a normal cat, as we have seen, many cortical cells receive fibers from both sets of geniculate-nucleus layers and therefore have a binocular input. These paired inputs are from geniculate cells near a single line of projection, which therefore represent the same point in the visual field, and there is an orderly point-to-point projection from the lateral geniculate nucleus to the cortex. The result is that the visual fields are mapped onto the cerebral cortex unambiguously and in an orderly sequence, so that as the cat moves in relation to its visual environment the moving outlines falling on the retina can be represented within the cortex as orderly waves of neural activation.

Let us now consider what might happen in the visual cortex of a Siamese cat, given its visual-pathway abnormality and assuming that the connections from the geniculate nucleus to the cortex were normal [*see illustration on page 52*]. First of all, some cortical loci would receive a contradictory input: neural activity at one cortical locus would represent visual stimuli in two disparate parts of the visual field, segments No. 6 and No. 11, for example. Furthermore, as the cat moved its head from one side to the other, instead of a single wave of activity passing from one cortical locus to the next there would be two waves, one representing a movement from, say, visual-field segment No. 11 to No. 12, as is normal, and the other representing a movement from, say, visual-field segment No. 6 to No. 7, which is abnormal. With such a system of connections it would be difficult for cortical activity to represent the visual world in an unambiguous way, and one is forced to wonder how a Siamese cat can see at all. Yet Siamese cats see quite well and are able to localize objects in space. How? Kaas and I studied cortical cells to determine how the visual fields are represented within the Siamese cat's cortex. Again we used microelectrodes and mapped the receptive fields of single cells or small groups of cells, essentially as we had earlier in the lateral geniculate nucleus.

The results of these experiments showed us that with a simple modification of the normal neural connections the cortex can still function as an unambiguous representation of the visual world. We found that in Siamese cats most cortical cells respond to stimuli coming through the normal geniculate-nucleus layers (Layer A in the illustrations), whereas the layers that are abnormally connected (Layer A_1 in the illustrations) have little effective visual input to the visual cortex; the cortical activity suggests that the cat is simply not seeing the input from Layer A_1 of each geniculate nucleus, which is to say from the temporal (outer) portion of each retina. There must therefore be a mechanism somewhere in the brain, probably in the cortex, perhaps involv-



VISUAL PATHWAYS are shown (*color*) in the cat's brain, seen from below and sectioned (*left*) to reveal the visual centers. Fibers from ganglion cells in the retina, bundled into the optic nerve, pass through the chiasm, where some of them cross to the opposite side of the brain. The fibers end in the two lateral geniculate nuclei. There they stimulate nerve cells whose fibers in turn project to the visual cortex of the left and right sides of the brain.



STAINED SECTIONS of the lateral geniculate nucleus of a normal cat (*left*) and of a Siamese cat (*right*) reveal the abnormality in the Siamese cat. In the normal brain the Nissl-stained nerve-cell bodies lie in two distinct layers designated A and A_1, as indicated in the map of the photomicrograph. In the Siamese cat, however, Layer A_1 is disrupted and parts of it appear to be fused with Layer A.

ing both the cortex and the lateral geniculate nucleus, that suppresses the whole of the abnormal sequence of inputs coming through Layer A_1. At present we know nothing about this cerebral censor. It remains to be determined what kind of neural circuitry can recognize the abnormal sequence and, having recognized it, suppress it.

While we were studying Siamese cats at Wisconsin several other groups were also examining this abnormal visual system. Quite recently we heard that Eva I. Elekessy and her colleagues at the Australian National University have been able to show through behavioral testing that Siamese cats act as if they cannot see with the temporal retina. When the cats are tested with both eyes open, they appear to be quite normal. It is only when one eye is blocked that an abnormality can be demonstrated: the cats react to visual stimuli on the same side as the open eye but not to stimuli that are beyond the midline. This is in sharp contrast to a normal cat, which can see about 45 degrees beyond the midline with the temporal retina of each eye. The behavioral evidence thus supports the view that in Siamese cats the input that would ordinarily come through Layer A_1 is suppressed.

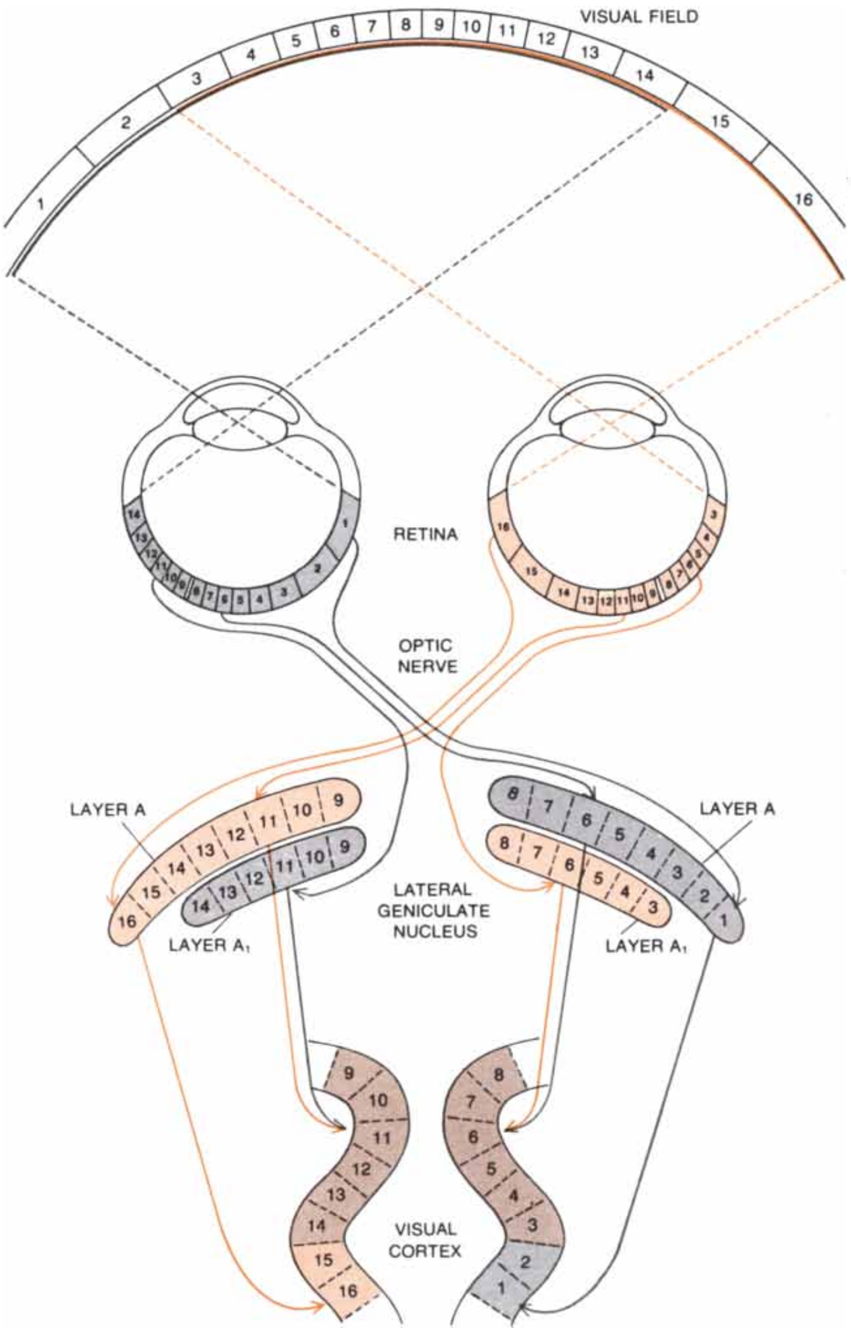
David H. Hubel and Torsten N. Wiesel had also been studying Siamese cats at the Harvard Medical School, recording from cortical cells with methods that did not differ markedly from ours. Disconcertingly, they obtained quite different results. In their cats, instead of the partial suppression there was a striking reorganization of the projection from the lateral geniculate nucleus to the cortex: the abnormal part of Layer A_1 was represented near the surface of the visual cortex. Not only was this representation of a part of the lateral geniculate nucleus inserted at an abnormal cortical site but also the sequence of the geniculate representation in the cortex was the reverse of normal. The rest of the lateral geniculate nucleus was represented in a normal sequence, but this normal representation was crowded by the insertion of the abnormal portion [*see illustration on page 53*]. The result is that the visual-field segments are represented in order (except for a slight break between the normal and the abnormal segments, between segments No. 8 and No. 10 in the illustration). In other words, the sequence that was disrupted in the geniculate nuclei has been put right in the cortex. It is difficult to avoid the conclusion that this dramatic rearrangement of the projection from the lateral geniculate nucleus to the cortex is established in order to compensate for the abnormal projection from the retina to the geniculate nucleus. It seems unlikely that the correction is the result of natural-selection pressures, since Siamese cats are bred for their appearance rather than for their visual abilities.

The pattern of cortical representation described by Hubel and Wiesel was so different from the pattern we had seen that we had to find some explanation for the difference. We could not reasonably blame different techniques, since their methods and ours were basically similar and the interpretation of the results is quite straightforward. To be quite certain of our results, however, we studied the projection from the geniculate nucleus to the cortex with another method: the method of retrograde degeneration. Since geniculate nerve cells send their fibers to the visual cortex in a well-organized point-to-point pattern, a small cortical lesion damages a well-defined group of geniculate nerve fibers and the nerve cells that give rise to these fibers show retrograde degenerative changes. In a cat with a normal projection from the lateral geniculate nucleus to the cortex a cortical lesion always produces a zone of retrograde degeneration that goes through all the geniculate-nucleus layers and is bounded by the lines of projection. The connections described by Hubel and Wiesel should lead to a quite different pattern of retrograde degeneration, with the zones in adjacent layers out of register [*see top illustration on page 54*].

We studied 11 Siamese cats with this method. The first 10 showed a normal pattern of degeneration. Finally, and fortunately, the last cat showed the pattern forecast on the basis of Hubel and Wiesel's results. We concluded that there must be two distinct patterns of projection from the lateral geniculate nucleus to the cortex in Siamese cats. One pattern, which we had encountered in most of our experiments, simply suppresses the abnormal inputs when they reach the cortex. We called it the Midwestern pattern. The other pattern does something more subtle: it re-creates an orderly representation of the outside world at cortical levels. We called it the Boston pattern. Both patterns allow the cortex to receive an unambiguous and sequentially organized representation of the visual fields and both use all the available cortical tissue. Neither, however, allows a significant amount of binocular interaction at cortical levels.

We do not know why some cat brains should solve their visual dilemma in one way and some in another. It may have to do with the size of the "medial normal segment," the portion of Layer A_1 that is not abnormal and is shown as segments No. 8 and No. 9 [*see illustration on page 52*]. The break between the normal and the abnormal representation on the cortex of Boston cats is related to the size of this medial normal segment. It is clear that the Boston pattern produces a complete sequential representation of the visual field only if the medial segment is absent or very small. The medial normal segment receives inputs from the central parts of the retina and is therefore likely to be particularly important for aligning the eyes. Cats in which the medial normal segment is smallest are likely to be the most cross-eyed; there is some evidence that severely cross-eyed cats show the Boston pattern, whereas straight-eyed cats tend to show the Midwestern one.

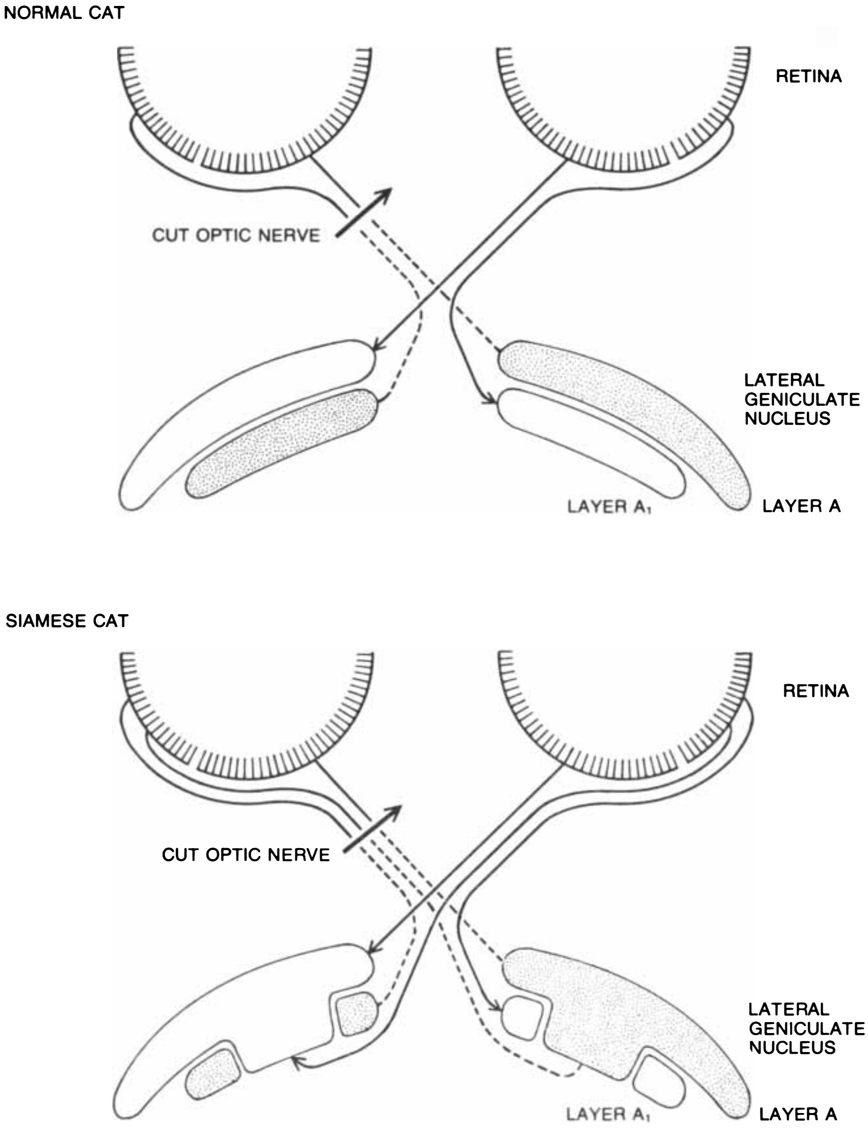
We still do not know why Siamese cats are cross-eyed; perhaps it is simply that there are very few cells in the Siamese cat's brain that receive a binocular input, so that the inputs from the two eyes never have to be matched. The information we have about the abnormality has, however, helped us to understand several features of the cat's central visual pathways. It shows that optic nerve fibers can be sent to the wrong side of the brain without any other modification of their highly specified pathways. The lateral geniculate nucleus accepts these fibers according to the locus of their retinal origin, and the standard topographic order of the geniculate-nucleus terminals is maintained even if its maintenance gives rise to a disrupted representation of the visual field within the lateral geniculate nucleus. The cerebral cortex, in contrast, will not accept a disrupted representation of the visual field. Instead the topographic order of the projection from the geniculate nucleus to the cortex can be altered quite



VISUAL FIELD is represented on the retina and in the brain of a normal cat as shown in this schematic diagram. The field is divided into 16 unequal segments. Each half of the brain "looks at" the half of the visual field on the opposite side and relatively more cerebral tissue is committed to representation of the central portions of the field than of the peripheral portions. An inverted image of the field is thrown on the retina by the reversing lens. The left eye does not see the extreme right side of the total field (segments Nos. 15 and 16) and the right eye does not see segments Nos. 1 and 2. Representations of the field that are seen by both eyes (segments No. 3 through No. 14) are localized in two sets of layers in the geniculate nucleus (*shown here as just two layers,* A *and* A_1), depending on whether they come from the eye on the same side or on the opposite side; the representations in the layers are in register, so that "lines of projection" (*broken lines*) through the layers represent single points in the visual field. The two sets of representations are fused in the cortex.

dramatically in some cats, and this reordering of the growth of the fibers recreates a sequential representation of the visual field. In some cats such a reordering appears not to be possible, and so the disrupted representation is suppressed. We have learned that the nature of the connections between the retina and the geniculate nucleus is quite different from that of the connections between the geniculate nucleus and the cortex, and we have been able to raise some puzzling questions about the mechanisms by which the central pathways are developed.

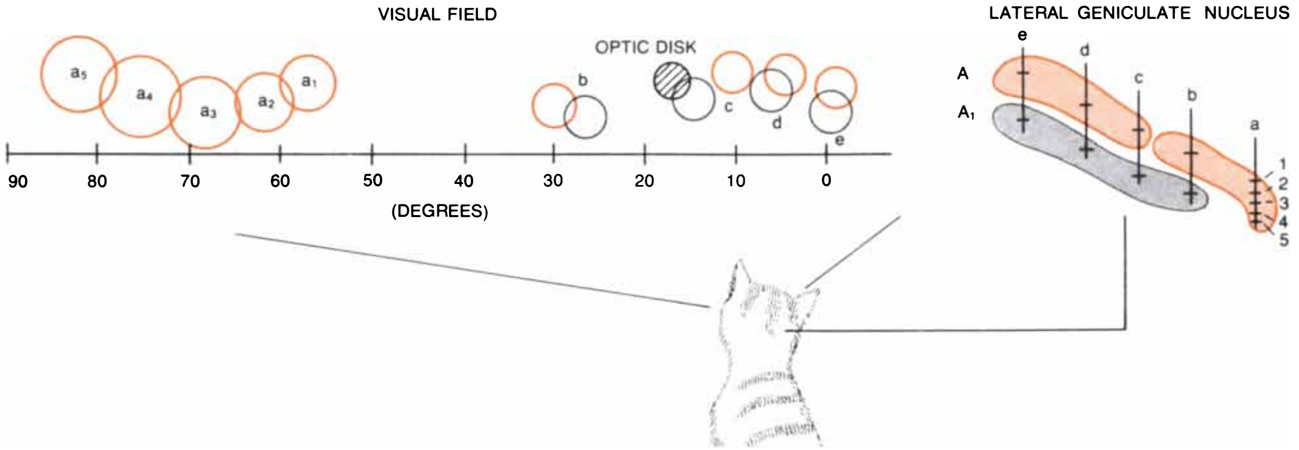
The Siamese cat is not a good animal in which to study prenatal development, and so it is fortunate that similar abnormalities have now been found in a number of other mammals. It was D. J. Creel of the University of Utah who first suggested that the aberrant pathways seen in Siamese cats might be an example of a much more widespread abnormality affecting albino animals in general. He pointed out that some years ago R. D. Lund of University College London had found fewer uncrossed nerve fibers coming from the retina in albino rats than in normally pigmented rats. Creel also drew attention to the fact that genetically Siamese cats represent a variant of albinism. Animals that have inherited an *albino* gene from both parents (that is, are homozygous for *albino*) are unable to make the pigment melanin at all because the *albino* gene cannot direct the synthesis of the enzyme tyrosine, which is essential for melanin formation. The gene *Himalayan*, which is a variant of *albino* that appears at the same locus on the chromosome, codes for an enzyme that can synthesize melanin at relatively low temperatures only, so that animals such as Siamese cats or Himalayan rabbits or mice, which are homozygous for the *Himalayan* gene, have white or pale bodies and darkened extremities.



SELECTIVE STAIN that is taken up preferentially by degenerating nerve fibers traces the path of fibers from one retina. One optic nerve is cut and its fibers degenerate. In a normal cat (*top*) the stain (*stippling*) shows that some fibers cross to innervate Layer *A* (upper layer) of the lateral geniculate nucleus on the opposite side of the brain from the cut; other fibers go to Layer A_1 (lower layer) on the same side as the cut. In the Siamese cat (*bottom*) there are extra fibers that cross to the opposite Layer A_1 and fewer that stay on the same side.

The suggestion that there is an "albino" abnormality of which the Siamese abnormality is only one example was of considerable interest. If the suggestion could be confirmed, there would be a number of different species in which the abnormality could be studied in more detail, and experimental approaches that are difficult in cats might prove to be easier in other species. If all albinos are abnormal, there is likely to be a simple genetic basis for the abnormality. And it would be particularly interesting to examine the extent to which the visual abnormalities that have long been recognized in human albinos might be related to abnormal visual pathways.

In order to demonstrate that there is a general albino abnormality it was necessary to show that the layers of the lateral geniculate nucleus in the albino rat, for example, are disrupted much as they are in the Siamese cat, and then to show similar abnormalities in other species. It turned out that the evidence about the rat had been available for several years but had seemed to represent merely a curious anomaly in the structure of the rat's lateral geniculate nucleus. The lines of projection, which represent single points in the visual field, had been defined for the rat by Vicente M. Montero and his colleagues at the University of Wisconsin Medical School on the basis of microelectrode studies. The lines run slightly obliquely through the nucleus [*see bottom illustration on page 54*]. When Montero and I studied the segments of retrograde degeneration that are produced in the rat's lateral geniculate nucleus by lesions in the visual cortex, the borders of these segments defined lines of projection that were in agreement with the electrophysiological results. When these lines of projection are compared with the cross sets of geniculate-nucleus layers that are defined by their retinal innervation, the lines should cross the layers, as they do in the cat and in many other species, if the representations of the visual field are to be in register within the lateral geniculate nucleus. W. R. Hayhow and



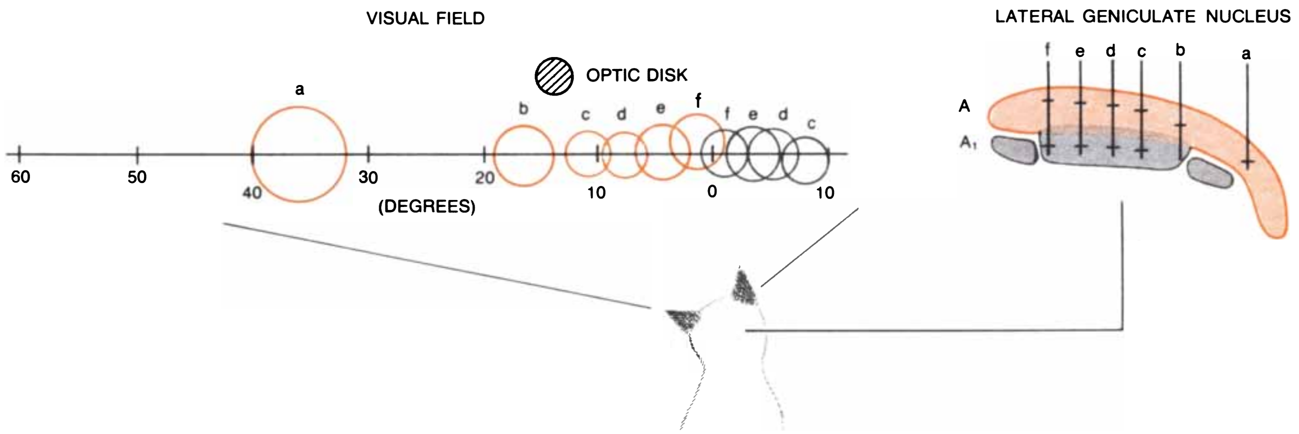
PROJECTION OF VISUAL FIELD on the lateral geniculate nucleus is determined by recording impulses in the geniculate nucleus (*right*) while stimulating various parts of the field (*left*). Results with a normal cat are shown here. Each lettered line represents a microelectrode puncture, with short horizontal lines indicating specific recording sites in the geniculate nucleus. The receptive field corresponding to each recording site—the location of a visual stimulus that elicited impulses at the site—is numbered correspondingly; the colored circles are fields for Layer *A* and the black circles are fields for Layer A_1. The optic disk is a blind spot at the point where nerve fibers leave the retina; note that there is a corresponding cell-free area between punctures *b* and *c* in the brain.

his colleagues at the University of Sydney had studied the rat's lateral geniculate nucleus, however, and had described layers that appeared to be oriented parallel to the lines of projection rather than perpendicular to them. A similarly anomalous situation existed for the rabbit.

We realized that these anomalous layers had probably been described on the basis of observations made in albino animals. That made it clear that they were likely not to represent complete layers at all but to be the interrupted pieces of an abnormal layer, much like the abnormal segments in Siamese cats. We were able to demonstrate that normally pigmented rats do show a continuous layer running across the lines of projection, just as one would expect. A similar analysis has recently been provided by Kenneth J. Sanderson for the anomalous layers in the rabbit.

We turned next to albinos of other species, looking for animals in which the lateral geniculate nucleus is clearly laminated, so that the abnormality might be readily recognizable, and in which albinism is relatively common. Ferrets, domesticated polecats that have been bred since Roman times for chasing rabbits out of their burrows, proved to be an ideal species, since albino ferrets are common and since carnivores generally have a well-laminated lateral geniculate nucleus.

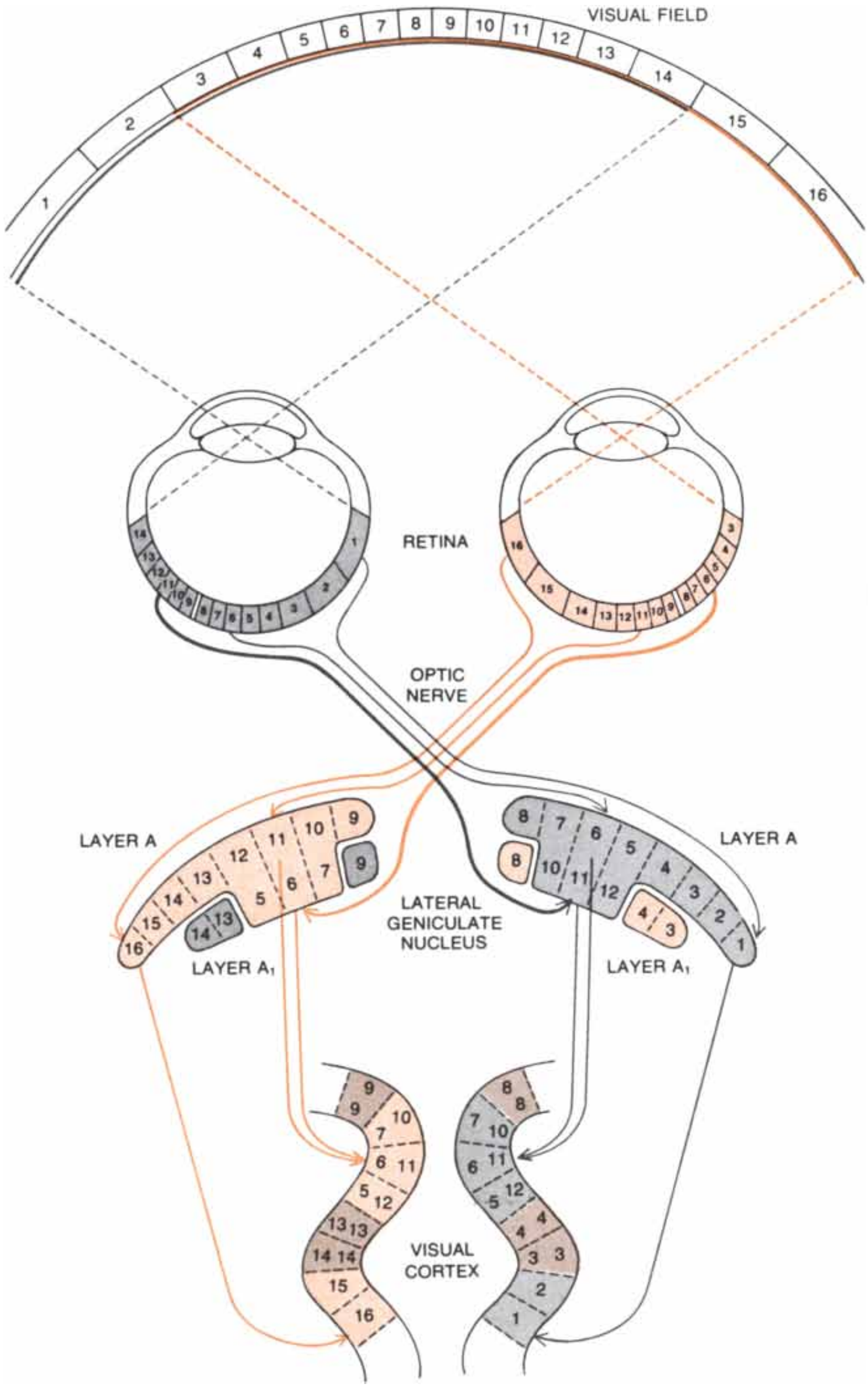
We found an abnormality in albino ferrets that was even more extensive than the abnormality seen in most Siamese cats. There are very few uncrossed fibers running from the retina to the lateral geniculate nucleus, and they end in two or three small geniculate cell groups; almost the entire projection from the retina crosses at the optic chiasm and terminates in anomalous segments of the geniculate-nucleus layers. Normally pigmented ferrets, on the other hand, have a set of layers that receive an uncrossed innervation. There was surprisingly poor differentiation of the visual pathways even in the normal animals. In comparison with the geniculate nuclei of other closely related carnivores the geniculate nuclei of normal ferrets are small and the layers are poorly defined. It is as though domestication had given rise to a reduction of the visual pathways, and one suspects that over the years the breeders of ferrets have tended to favor



PROJECTION OF FIELD in Siamese cats was similarly determined. The representation of the field in Layer *A* (*color*) is normal; as in the normal cat, sites ranged from the outer to the central part of the right lateral geniculate nucleus have receptive fields ranging from the outer to the central parts of the left visual field. In Layer A_1, however, the representation of the visual field is abnormal (*black circles*): it undergoes a mirror reversal. (In this case the electrode punctures missed the site at which the optic disk is represented in the geniculate nucleus, so that no cellular discontinuity is seen and the receptive fields are below the blind spot.)

ferrets with poor vision; the suspicion is strengthened by the fact that the albino ferrets, which are known to have poor vision, have been strongly favored.

Recently albinos of several other species have been studied, including albino mice, albino guinea pigs and albino mink, and they all show the same basic abnormality. So do Himalayan rabbits and hamsters. One animal that was of unusual interest was the so-called white tiger. Robert W. Doty of the University of Rochester School of Medicine and Dentistry pointed out to us that one of these tigers, a female named Rewati at the National Zoological Park in Washington, is cross-eyed. (Rewati is temporarily at the Brookfield Zoo in Chicago.) We were naturally not able to examine this animal's brain, but we were able to study the brain of a brother named Moni that had died some years ago. Moni's brain had been preserved in the Armed Forces Institute of Pathology, and when Kaas and I examined it, we found a disruption of the geniculate-nucleus layers. The disruption is milder than what is seen in Siamese cats or albino ferrets but the laminar pattern was clearly abnormal in relation to the pattern found in normally pigmented carnivores. We concluded that the central visual pathways of the white tiger are also abnormal.



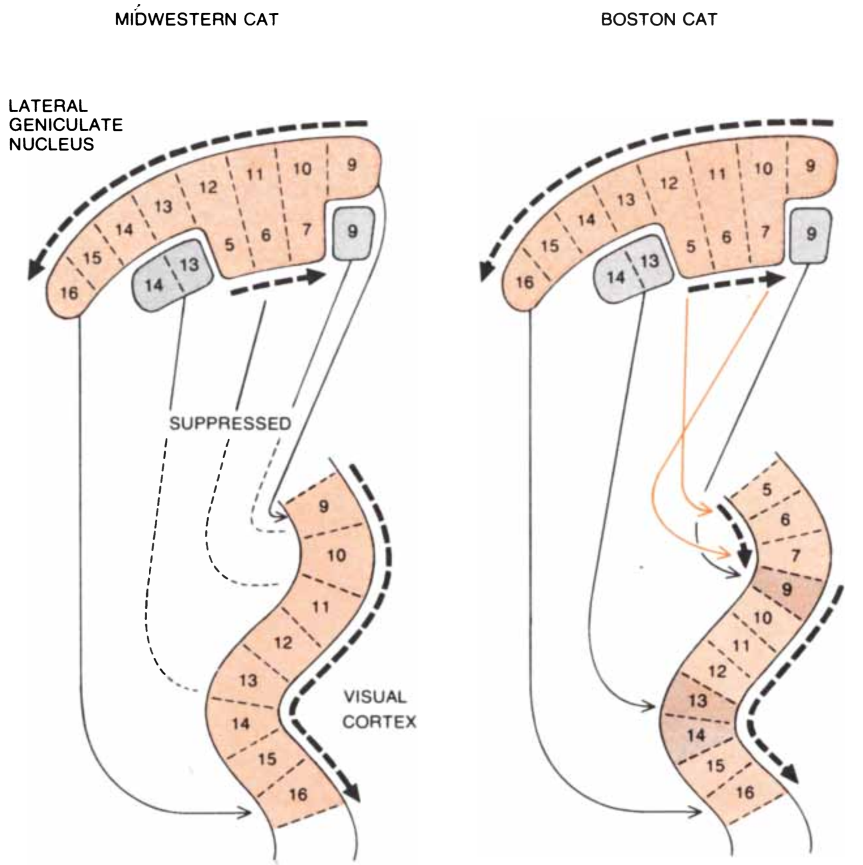
ABNORMAL PATHWAYS (*heavy arrows*) to the lateral geniculate nucleus would have serious effects on the representation of the visual field in the cortex, if the connection from the lateral geniculate nucleus to the cortex were normal. Projections from the two layers of the lateral geniculate nucleus, instead of reinforcing each other, would in many cases conflict. Neural activity at a single point in the cortex would have an ambiguous meaning, reflecting stimuli in two parts of the visual field, such as segments Nos. 6 and 11.

The white tigers are not really white. They have normal gray-brown stripes, but in place of the usual orange stripes they have white ones. Their eyes are blue, rather like those of a Siamese cat, but there is very little pigment in the pigment epithelium, the layer behind the sensory part of the retina. This partially reduced pigmentation resembles the effect of the *chinchilla* gene, another member of the albino series. Chinchilla rabbits and Burmese cats show a comparable pigment deficiency, but we have been unable to find any abnormality of the central nervous system in these animals. It is probably relevant that the pigment epithelium in chinchilla rabbits is quite normal. It would seem that the *chinchilla* gene gives rise to either a mild abnormality or no abnormality at all, whereas the other two genes we have studied, *albino* and *Himalayan,* produce greater pigment deficits and also more extensive abnormalities of the visual pathway.

The relation between the pigment deficit and the abnormality of the visual pathway is intriguing. The abnormality involves the retinal ganglion cells, some of which must be wrongly specified with respect to the course of their fibers toward the brain. A simple mechanical deviation of a bundle of fibers at the optic chiasm appears to be ruled out because that would not produce abnormally connected retinal patches with the rather sharp borders we have found; it looks instead as if certain retinal cells are somehow specified for incorrect routing. The problem is that retinal ganglion cells do not make pigment, and one has to ask whether they can be influenced by pigment-forming cells in neighboring parts of the developing retina. Alternatively, the ganglion cells may have some intrinsic mechanism that depends on the enzyme tyrosinase without intracellular pigment production.

We do not know how the albino genes act, but we have recently tried some indirect approaches that may help us to define the mechanism. One method has been to study mice that have some albino cells and some normally pigmented cells. These "flecked" mice are females in which a piece of the chromosome that normally carries the *albino* gene is attached to an *X* chromosome. Mary F. Lyon of the Medical Research Council unit at Harwell in England had pointed out that since females have two *X* chromosomes and males only one, in normal females the genes on one of the *X* chromosomes are inactivated and there is only a single active gene at each gene locus in both sexes. The inactivation appears to be random: in some cells one of the two *X* chromosomes is inactive and in some cells the other is. In the flecked mice, which were reared by Bruce M. Cattanach at Harwell, there is only one wild-type (normal) variant of the gene, which forms pigment, in each cell, and that gene is on one of the *X* chromosomes. In addition there are two *albino* genes at the normal locus. Some cells, in which the wild-type gene is on the active *X* chromosome, are pigmented; other cells, in which the *X* chromosome with that gene is inactive, are albino. Malkiat S. Deol of University College London and Wesley K. Whitten of the Jackson Laboratory in Bar Harbor, Me., have shown that in the flecked mice the pigment epithelium is made up of patches, some normally pigmented and some albino.

Cattanach, Deol, Grayson L. Scott and I decided to study the albino abnormality in these flecked mice. We argued that the size of the abnormality, if it occurred in flecked mice, might tell us more about the mechanism by which the course of the fibers from the retina to the lateral geniculate nucleus is specified. If each retinal ganglion cell acts autonomously in this specification, so that the mechanism is essentially intracellular, then the flecked mice should show an abnormality about half as large as albino mice do (that is, they should have more fibers running a normal course to the same side of the brain), since only about half of the cells lying in the potentially abnormal patch of retina are albino. We found that the abnormality in the flecked mice is not half the size of the albino abnormality, which indicates that the pathways are not specified by an autonomous intracellular mechanism. (Curiously, we found that the flecked mice actually had more fibers going to the same side than normal mice, and we are now trying to find out more about this odd result.) The only conclusion we can draw from our observations so far is that the mechanism specifying

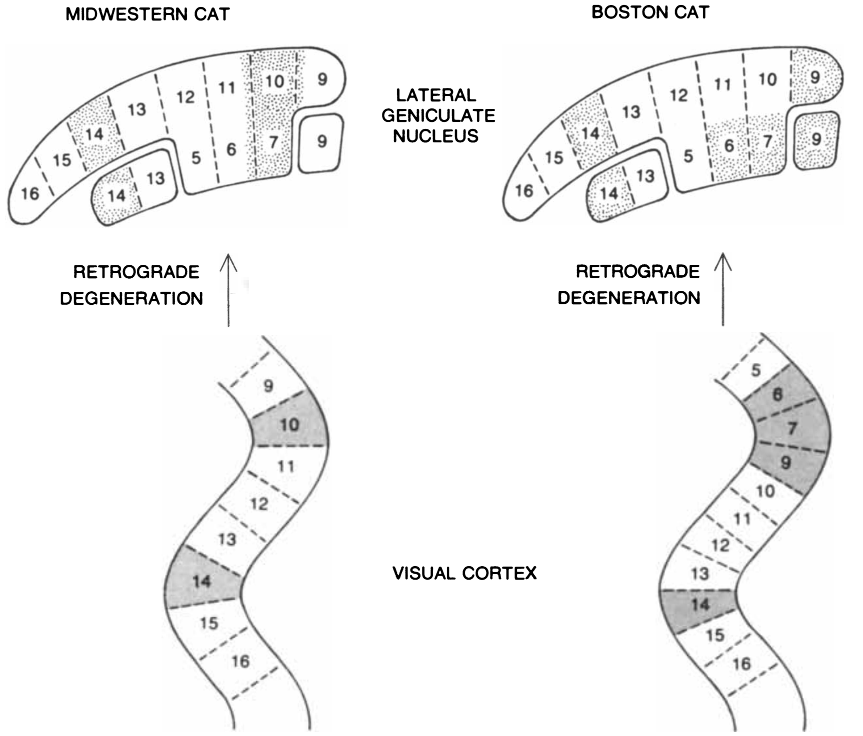


COMPENSATION for the abnormality is achieved in one way in the "Midwestern" Siamese cats and in another way in "Boston" Siamese cats, as illustrated here for the left side of the brain. In a Midwestern cat the input to the cortex from the A_1 layer is largely suppressed, eliminating the misdirected representation of segments Nos. 5, 6 and 7. In a Boston cat the geniculate-nucleus representation of segments Nos. 5, 6 and 7 is instead reversed by a rerouting of the fibers to the cortex (*colored arrows*) and is inserted at the surface of the visual cortex. The misdirected part of the visual field, although projected to the wrong cortex, thus "reads" correctly as a more or less continuous picture of the outside world.

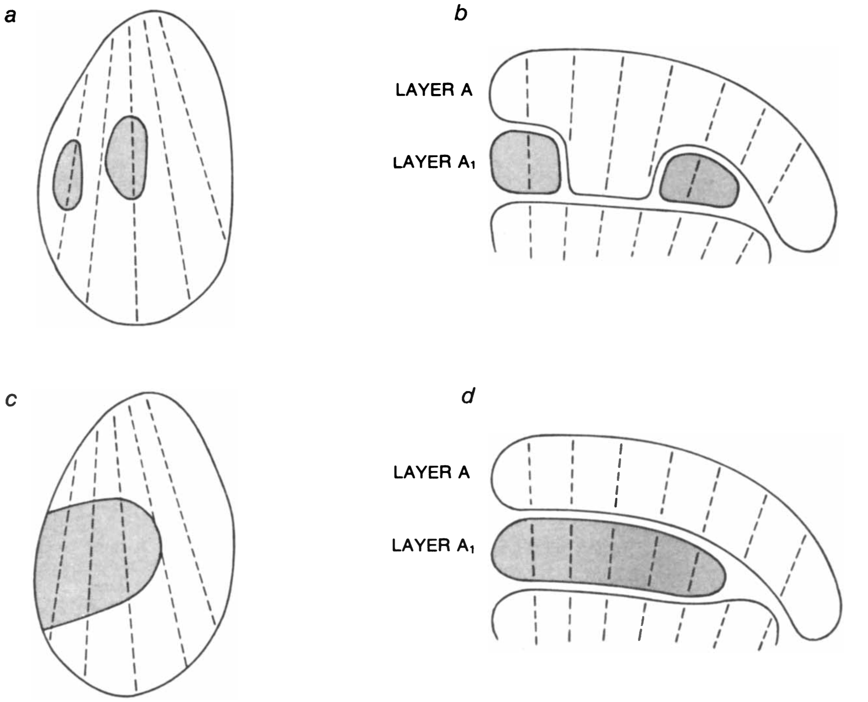
the pathways is likely to involve interactions among cells within the retina. That is, although the *albino* gene acts through an intracellular mechanism to produce pigment, it appears to act through an intercellular mechanism in controlling the visual pathways.

Further information about the genetic basis of the abnormality of the visual pathways has come from studies of ranch mink. Mink breeders put a high value on white mink, and so a variety of gene combinations that can produce a pigment deficit have been preserved in this animal. The mink has good binocular vision and its geniculate-nucleus layers are well defined. Abnormalities of the visual pathway are therefore easier to study in the mink than in species, such as the mouse or rat, in which the layers are poorly defined. Sanderson, Richard Shackelford and I have studied several genetically distinct mutations in the mink and have found that all gene combinations that give rise to a significant lack of retinal pigment also produce abnormal visual pathways. The extent of the abnormality is inversely related to the amount of retinal pigment; white animals with normally pigmented eyes show no abnormality of the central visual pathways.

Two important conclusions arise from these observations of mink. The first is that the *albino* gene is not unique in producing the abnormality of the visual pathways. Gene loci quite distinct from the *albino* locus can produce abnormal visual pathways, at least in the mink, and so we cannot regard the lack of tyrosinase or any other single enzyme as being uniquely responsible for the abnormality. The second conclusion is that retinal-pigment formation itself is closely related to the development of the abnormality. The details of this relation are not known, but it is interesting that pigment formation in the retina precedes the differentiation of the retinal ganglion cells, and that the final cell division giving rise to those cells takes place in a



RETROGRADE DEGENERATION of cell bodies in the lateral geniculate nucleus (*black dots*) after fiber endings in the cortex have been damaged (*gray areas*) reveals the paths of fibers from the lateral geniculate nucleus to the cortex. As expected, the paths differ in Midwestern and Boston cats. The "normal" connections from the lateral geniculate nucleus to the cortex in Midwestern cats produce zones of degeneration that go through both layers. The abnormal (but corrective) connections to the cortex in Boston cats produce zones of degeneration in the layers of the lateral geniculate nucleus that are out of register.



ALBINO RAT has been shown to have a geniculate-nucleus abnormality comparable to the Siamese cat's. Microelectrode and retrograde-degeneration studies had traced lines of projection (*broken lines*) in the rat (*a*) that appeared not to cross different layers; the layers seemed to be "islands" (*gray*) oriented parallel to the lines instead of across them. The experimental rats were albinos, however; the "islands" are really abnormal interrupted segments of layers like those in Siamese cats (*b*). Normally pigmented rats turn out to have continuous cell layers (*c*) that are actually quite analogous to the layers in normal cats (*d*).

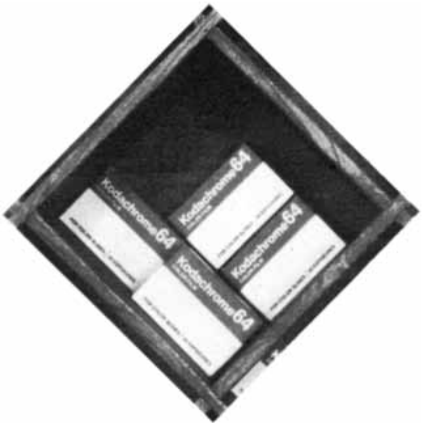
layer adjacent to the pigment epithelium. At about the time that many of the ganglion cells are forming, the embryonic eye cup is also undergoing a major modification: a large slit called the fetal fissure, through which the retinal artery enters the eye, is in the process of closing. It is an odd fact that the region in which this closure occurs is entirely free of pigment. Exactly how the distribution of pigment, the closure of the fetal fissure and the specification of the ganglion cells are related to one another, if they are related at all, remains to be determined.

Eventually, of course, one wants to know about human albinos. Do they also have the abnormal pathways? If they do, how does that affect their vision? Albinism in man is generally associated with very low visual acuity and with the rhythmic horizontal movement of the eyes called nystagmus. There is often a misalignment of the eyes too, so that the visual responses and abilities of albino people are often difficult to study. Examination of brains after death might reveal an abnormality, as our examination of the tiger brain did, but so far no such reports of human albino brains are available.

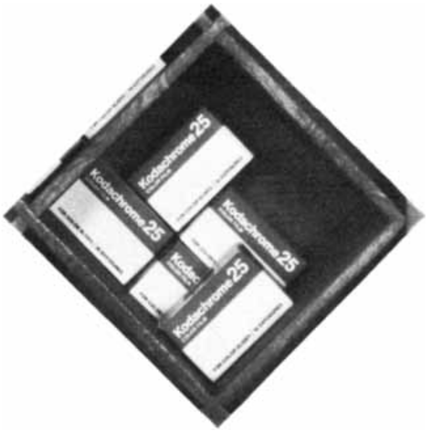
Further investigations of the albino abnormality can lead in several directions. The visual abilities and central visual connections of human albinos are of obvious interest, although they may prove difficult to examine. Experimental studies of embryonic development in albino animals can be expected to show the mechanism by which genes act to modify the development of brain pathways. An understanding of how one gene acts in cerebral development should help us to determine the mode of action of many other genes that play a part in neural development. Perhaps the most intriguing aspect of the albino story is that in some brains the central connections can be modified to correct a scrambled input. If we could find out more about the developmental processes by which the sensory representations are thus put right, we might be well on the way to understanding how the brain normally acquires an orderly representation of the outside world.

The genetic abnormality can be regarded as a delicate experiment that could not have been set up by any surgical means. On the one hand it enables us to study developmental processes that give rise to cerebral abnormalities. On the other hand it affords us an opportunity to determine the extent to which central pathways can adapt to deal with abnormal patterns of cerebral activity.

Three new Kodak products

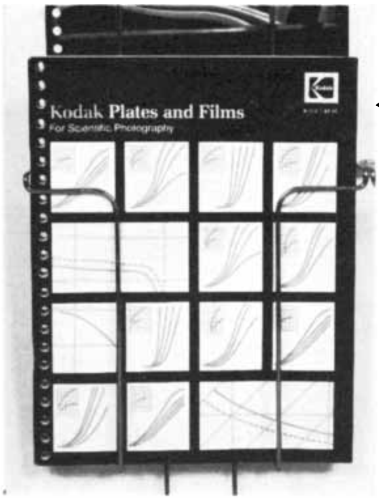


KODACHROME 64 Film (Daylight) in 135, 126, and 110 size, for color slides. Same exposure index 64 as Kodachrome-X film, which it will replace in the interests of cleaner whites, cleaner yellows, brighter reds, brighter greens, brighter blues, a better blue sky, more exposure latitude, more sharpness, less graininess. Fails, as do all known photographic processes, to duplicate *exactly* all color sensations aroused by the real world.* About as sharp as Kodachrome II professional film (type A), but not quite as sharp as:



KODACHROME 25 Film (Daylight) in 135 and 828 size, for color slides. Successor to daylight-type Kodachrome II film. Same exposure index 25 as its predecessor, with more vivid color rendition but same extremely fine granularity and high sharpness. (Book advertised below explains the difference between those terms.) Quality in color slides from a given camera depends on (1) the film manufacturer, (2) you, (3) the film processor. Choice of (3) and (1) is open. To help with factor (2), literature is available for all levels of interest. In the processing of this film, (1) gives (3) a better chance for good control.

These new films may not be obtainable everywhere yet. Their predecessors aren't bad.



← $2.50

If this new Kodak publication is not in the literature rack at your photo retailer's, it may be that the retailer doubts it will sell like Kodachrome film. If you'll designate the title and "No. P-315" and send payment to Kodak, Dept. 454, Rochester, N.Y. 14650, we can mail you a copy.

It is descended over many generations from a pamphlet by the founder of the Kodak Research Laboratories. As a young man in England with a fresh doctorate, C.E.K. Mees landed a job with Wratten & Wainwright, makers of dry plates, and immediately proceeded to bring science to photography and photography to science. This called for some expository writing. Soon demand for Mees's pamphlets by the technically inclined became comparable to their demand for the products themselves. Whereupon a price was put on them and they became products.

With "Wratten" still alive and well as a Kodak trademark, we continue to feel that in the photographic field, concise and authoritative information on the nature and applications of the products can be a valuable commodity in its own right.

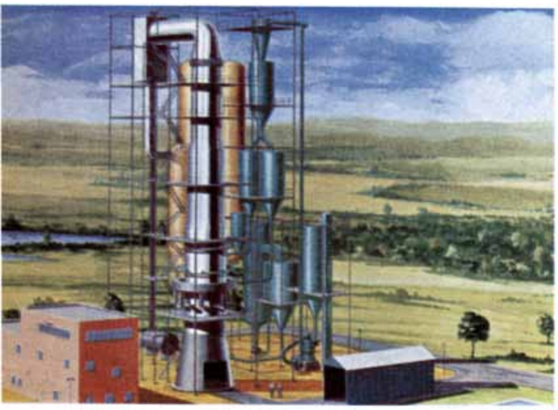
True, the book ignores color photography and contains data on only one manufacturer's products. More valuable than the data, perhaps, are the pages that explain derivation and use of photographic speeds, techniques to improve detection of weak signals, modulation-transfer characteristics, image effects that sometimes mislead, and many other such matters that users of the silver halide emulsion may not wish to study in depth but may find hard to pick up casually in the cafeteria.

The book winds up with a suggested course of action entitled "How to Select a Photographic Material for Special Applications." It helps weigh the relative advantages of materials stocked by nearby suppliers as against factory-stocked items and those that may take us up to four months to turn out for you.

Kodak

*More on that, if you like, in R. W. G. Hunt's paper "Objectives in Colour Reproduction" (*The Journal of Photographic Science 18*:205). (Reprints previously offered here and still available from Dept. 55W, Kodak, Rochester, N.Y. 14650.)

Taking coal out of the Dark Ages.



Now we can "mine" for gas. By 1985, some of the gas we use will be made from coal. Two C-E divisions are now underway on a program being funded by the Office of Coal Research that will lead to construction of a 200 MW pilot gasification plant. The "entrainment" process used will ultimately produce low Btu, clean burning gas for utilities . . . helping ease demand on more critical petroleum fuels.

You know how plentiful oil is these days.

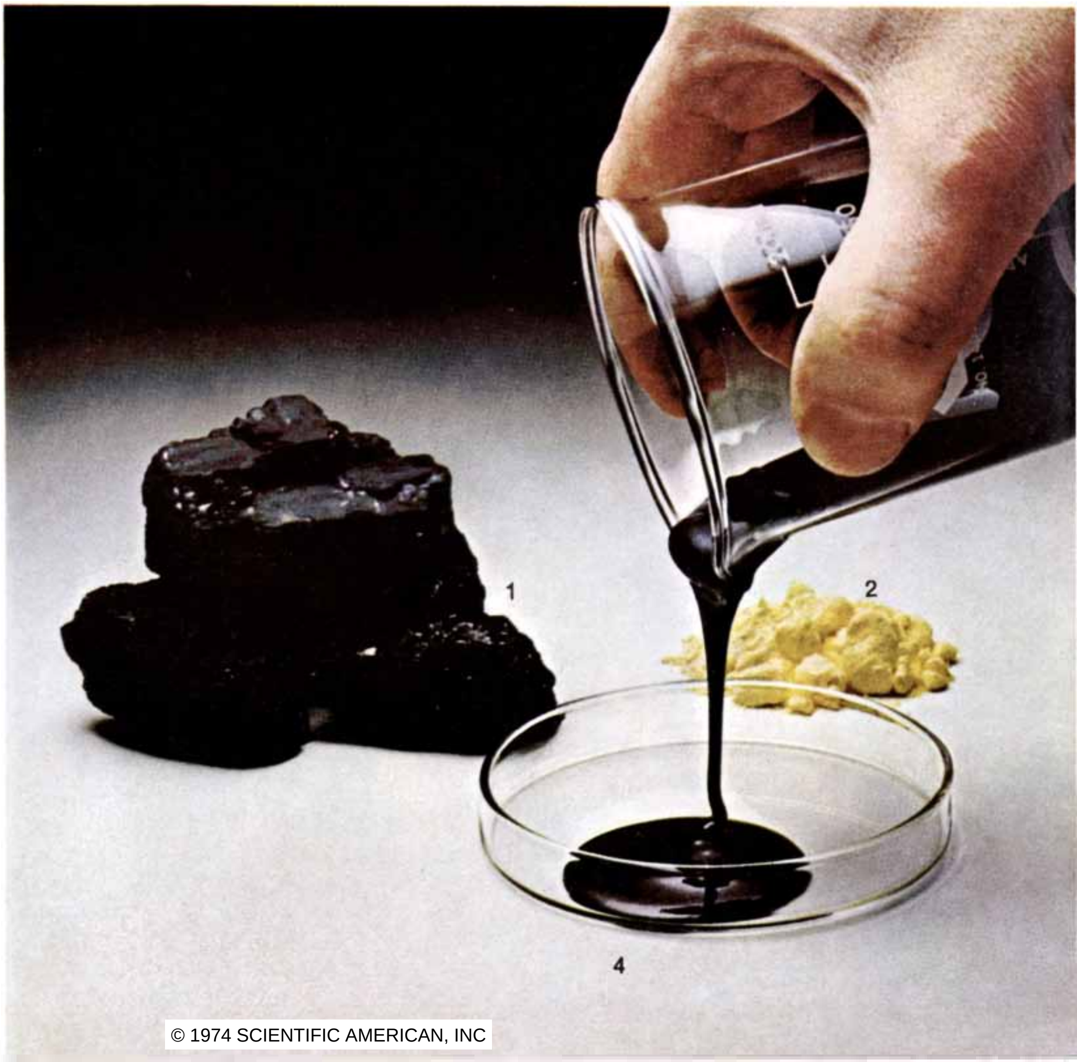
Did you know we can make "oil" from coal? And that we can also manufacture synthetic gas from this same, abundant energy source?

America owns 40% of the world's proven coal reserves.

Why haven't we used more? Three major factors are:

Environment: most easily recoverable coal does not burn "clean." Desulfured coal and improved anti-pollution devices will ultimately provide the answers. Modification of clean air standards should help coal provide emergency relief during the interim.

Economics: price controls have made opening new mines unattractive . . . and have diverted much of our coal to more lucrative for-



Clean-burning liquid fuel from high sulfur coal. East of the Mississippi, much of our coal (1) contains 3-5 times too much sulfur to meet EPA standards. C-E is perfecting a solvent refining process that will remove roughly half an ounce of sulfur (2) and 1½ ounces of ash (3) from a typical pound of coal treated. End product: a clean burning liquid (4) that can be used by industry and electric utilities. Clean fuels will have their price. Turning some high sulfur coals into liquid, for example, eats up 30 to 35% of the resource's original energy content.

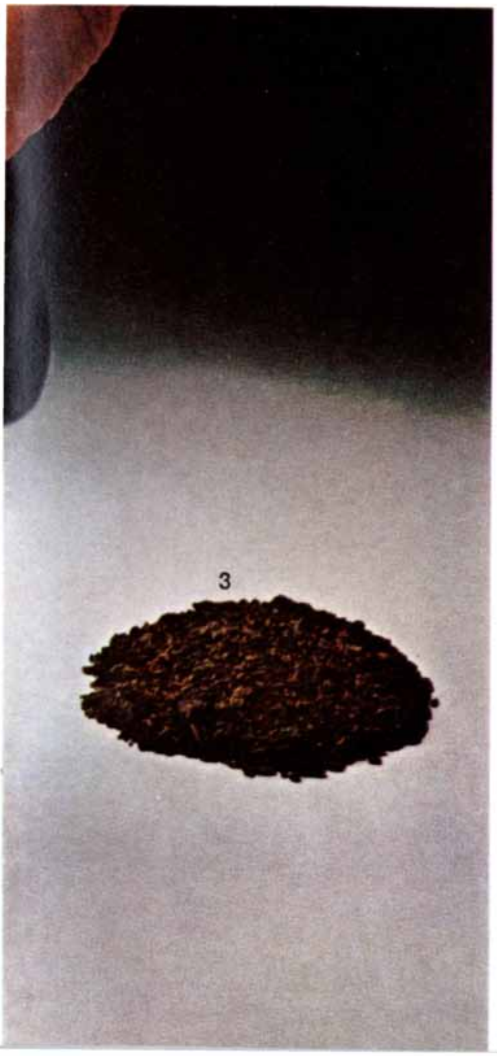
Burning dirty, but smelling sweet. Over 50 technologies for reducing sulfur in smokestack emissions are now being evaluated. One promising C-E system is being refined to improve already impressive operating efficiency and availability. When perfected, power plants will burn high sulfur coal and still meet EPA standards. But again, clean air won't come cheap: a system for *one* existing 500 MW power plant will cost about $25,000,000 . . . and reduce salable electric output some four percent.

eign markets. If we want more coal, we must be ready to pay reasonable, competitive prices.

Research and development: perfecting new coal technologies will require massive research investment. Expenditures for new facilities must also be sharply increased. "Liquid coal" for a single large power plant, for instance, needs a facility costing $300,000,000. To meet these challenges, greater support from public as well as private sectors is mandatory.

By 1985, there will be a demand for 2 to 3 times the coal we now extract. Combustion Engineering is helping coal point the nation toward energy independence. With promising technologies in coal liquefaction and gasification. With improved stack emission control devices. And as the world leader in coal-fired, steam generating systems.

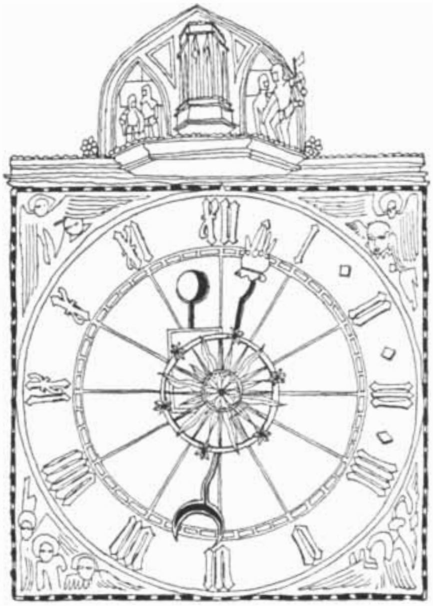
For more information on today's energy challenge, please write Combustion Engineering, Inc., Dept. B., 900 Long Ridge Road, Stamford, Connecticut 06902.



Strip mines can be made beautiful. *Top:* even as the coal is being stripped from the pit, reclamation gets underway. *Bottom:* Mine Lake, Illinois, was once a strip mine. While it may be desirable to curtail stripping in mountainous regions, we have abundant coal deposits in open, reclaimable areas. But preserving the land will contribute to coal and electricity price rises: reclaiming one acre can easily cost ten times its purchase price.

COMBUSTION ENGINEERING, INC.

THE ENERGY SYSTEMS COMPANY



SCIENCE AND THE CITIZEN

Sidereal Messengers

The past five months have been a time of splendor for planetary astronomy. First *Pioneer 10* cruised past Jupiter on December 3, photographing the planet and its red spot in unprecedented detail and finding an atmosphere around its largest satellite, Io. *Pioneer 11,* launched a little more than a year ago, emerged from the asteroid belt on March 20 and is now three-fourths of the way to Jupiter; its course has been altered so that it will also swing around Saturn in September, 1979. On February 5 *Mariner 10* flew past Venus, sending back its extraordinary pictures showing the dynamic architecture of the planet's atmosphere. Then on March 29 the same spacecraft transmitted revelatory information about the surface, history and magnetism of Mercury.

The *Mariner 10* pictures showed that Mercury strongly resembles the earth's moon in that its surface is much cratered, little eroded and has large, smooth areas reminiscent of the moon's maria, or "seas." Moreover, data on the infrared radiation from the planet's dark side and illuminated side indicate that its surface material is much like the moon's: low in density and therefore probably rich in silicon and poor in iron. The overall mass of Mercury, however, determined with unprecedented accuracy from the trajectory of the spacecraft as it moved past the planet, tells a different story.

The mass figure indicates that the overall density of Mercury is 5.5 grams per cubic centimeter, compared with 3.3 grams for the moon. Since the surface of the planet appears to be composed of light material, its core must therefore consist of heavy material. It is probably rich in iron and poor in silicon. Thus Mercury, unlike the moon and like the earth, is a highly differentiated body.

Yet the cratered surface of Mercury indicates that we are seeing it as it was soon after its accretion from many smaller bodies. There is no indication of tectonic processes of the kind that may have caused the earth's lighter materials to rise to the surface and its heavier materials to sink toward the core. The inference is that the differentiation of Mercury took place before the last stages of its accretion. If this is so, general concepts of the formation of the planets will have to be revised.

An iron core would help explain surprising data sent back by the magnetometer on *Mariner 10*. It had been expected that Mercury would not show a magnetic field, because it rotates slowly and does not emit radio waves. The latter fact had implied that the planet has no belts of charged particles, further implying that it has no magnetic field capable of trapping such particles. Every expectation was that the solar wind (ionized gases from the sun) would be observed to flow smoothly past Mercury as it does past the moon, leaving a cavity on the side of the planet facing away from the sun. The results were to the contrary. The magnetometer detected the presence of a bow shock wave that can only result from the deflection of the solar wind by an influence somewhat larger than the planet itself. Preliminary analyses suggest that the magnetic field at Mercury's surface has a strength of 200 to 300 gammas. (The magnetic field of the earth at the Equator has a strength of 30,000 gammas.)

Unruly Matter

One of the most successful inventions of theoretical physics in recent years has been the idea that the subtle order apparent in the relations among the "fundamental" particles of matter can be accounted for in terms of some even more basic, still undetected constituents: fractionally charged entities (called quarks by some theorists and partons by others) that join in various combinations to form the plethora of particles discovered in high-energy accelerators and cosmic rays. This approach has been particularly helpful in explaining the behavior of the large class of particles, called hadrons, that "feel" the strong nuclear force.

It therefore comes as a surprise that some persistent new results emerging from experiments involving colliding beams of electrons and positrons (antielectrons) cannot be explained by either the quark model or the parton model (or indeed by any other current model of the underlying nature of matter). The latest experimental findings, made at SPEAR, the electron-positron storage ring at the Stanford Linear Accelerator Center, confirm that hadrons are produced much more frequently than the theoretical models had predicted in electron-positron collisions at high energies. This unexpected overproduction of hadrons had been observed previously at the ADONE electron-positron storage ring at Frascati in Italy and at the Cambridge Electron Accelerator in the U.S.

According to the quark model, the total "reaction cross section" for the production of hadrons in the mutually annihilating collision of an electron and a positron should fall off as the square of the energy; in contradiction, the total cross section for hadron production measured in the SPEAR experiments remains virtually constant at energies of up to 5 GeV (gigavolts, or billions of electron volts). Moreover, in all the fractional-charge models the ratio of the cross section for producing hadrons to the cross section for producing pairs of muons is expected to remain constant at such high energies; instead this ratio is found to increase as the center-of-mass energy of the colliding particles approaches 5 GeV.

Perhaps even more surprising to theorists is the failure of the new results to conform to the concept of "scaling": the expectation, supported by earlier experiments, that the energy distribution of the hadrons emerging from the electron-positron collisions would exhibit a certain regularity. The scaling concept, which would enable one to predict the energy distribution of the hadrons at any energy once they had been measured at another energy, is currently a feature of every major model of particle behavior, including the quark and parton models.

It has been noted, in fact, that the energy distribution of the hadrons emerging from electron-positron collisions at high energies resembles nothing so much as the energy distribution of the particles emerging from very-high-energy proton-proton collisions. This observation has in turn suggested to some that the electron may be sensitive to the strong nuclear interaction within a radius of 10^{-16} centimeter. If this is true, it would appear that the electron, the best-known of the leptons—particles that supposedly do not feel the strong force—spends part of its life as a hadron.

The Quasar as a Galaxy

Quasars, the ultraluminous starlike objects that seem to qualify as the most remote bodies in the universe, have perplexed astronomers since their discovery some 13 years ago. Their inferred distance is based on the large red shifts in their spectra, which indicate they are receding at extremely high velocity. According to current cosmological concepts, the higher the recession velocity, the more distant the object. If quasars are as distant as they seem, they must emit 50 to 100 times more energy than entire galaxies consisting of hundreds of billions of stars. There is no satisfactory explanation for how an object as small as a quasar (at most a few light-years across compared with 100,000 light-years for a galaxy) can be so luminous.

According to one hypothesis, quasars are not a special class of objects but simply the visible manifestation of a titanic explosion in the nucleus of a galaxy otherwise as unremarkable, perhaps, as our own (see "The Center of the Galaxy," by R. H. Sanders and G. T. Wrixon; SCIENTIFIC AMERICAN, April). If the explosion were to take place in a galaxy so distant that it would ordinarily be invisible, all one would see on photographic plates would be the intense emission of the explosion itself. Actually a few quasars have seemed to be embedded in a fuzzy structure that might represent a faint surrounding galaxy, but this has remained to be demonstrated.

The first direct confirmation that a quasar is associated with an otherwise normal galaxy has now been reported in *The Astrophysical Journal* by J. B. Oke and James E. Gunn of the Hale Observatories and the California Institute of Technology. Oke and Gunn's discovery came after four years of study of BL Lacertae, which is listed in astronomical atlases simply as a variable star, presumably one of millions within our own galaxy. Their attention was drawn to BL Lacertae a few years ago when it was observed that the object also emits strongly in the radio part of the spectrum. On close examination BL Lacertae seemed to consist of a bright core within a fuzzy ring. The core also fluctuated rapidly in brightness, often from night to night. Thus it resembled a quasar in certain important respects. Unfortunately all attempts to find spectral lines in the light from BL Lacertae failed; hence there were no clues to its possible distance. If it were simply a peculiar radio-emitting object inside our own galaxy, it would not qualify as a quasar.

Undeterred, Oke and Gunn made a fresh effort to record a spectrum using a special obscuring disk that, when installed in the 200-inch Hale telescope, blocked out BL Lacertae's bright central core but not the fuzzy ring around it. (The disk was designed and built by a colleague, Earle B. Emery.) With the aid of the disk, Oke and Gunn were able to record the spectra of the material that composed the fuzzy halo.

It was quickly evident that the halo was composed of typical older stars of the kind found in giant spherical galaxies. Moreover, the spectra exhibited a red shift indicating the galaxy was about a billion light-years away. With the distance established, the bright central core of BL Lacertae could confidently be classified as a quasar—the nearest quasar yet discovered. The fact that the observers could not detect a spectrum of the core may be owing to the absence of heated gas clouds, which account for the spectra of most quasars. Normal spherical galaxies characteristically contain very little gas. Oke and Gunn's work will make it harder for skeptics to question the enormous distances and luminosities assigned to other quasars.

Good Behavior

Behavior modification is the application of psychological principles and techniques in an effort to eliminate what is considered socially undesirable and promote acceptable behavior. Proponents say it promises to control criminal activity more effectively than present correction methods and more humanely than any prison system can. Opponents say it is an infringement of human and constitutional rights, may involve dangerous tampering with personality and is often simply punishment by a fancy name. Recently a behavior-modification project at a Federal institution in Missouri was suspended after a court attack and the Federal Law Enforcement Assistance Administration announced that it would no longer apply anticrime funds to behavior-modification programs and experiments. The two actions did nothing to still the basic controversy.

One problem is that "behavior modification" covers a wide range of approaches. At one extreme it merely involves attempts to shape behavior by "reinforcing" what is desired, as has long been done in animal experimentation and in some human psychotherapy. At the other extreme it contemplates the monitoring and control of an individual's actions by means of telemetry and brain stimulation. The electronic procedures have been tested with volunteers but are far from being put into practice, and even psychologists who favor the possibility assume that stringent safeguards would be necessary. Reinforcement through the manipulation of "rewards," on the other hand, has been utilized in some institutions for a number of years. Critics of these methods, such as the Children's Defense Fund of Washington, point to a number of abuses. For one thing, no inmate of an institution can really give effective consent to any program in which he is involved. Most programs, however grandly conceived or psychologically described, are administered by prison or mental-hospital guards or attendants. Finally, the manipulation of rewards usually involves not real reward but unusual punishment; a standard technique is to deprive a person of something with the promise that good behavior will end the excessive punishment.

Cancer on the Caspian

Among the least common but most fatal forms of cancer is one that begins as a tumor of the cells lining the esophagus, the section of the alimentary tract between the pharynx and the stomach. In two sizable districts bordering on the Caspian Sea, however, cancer of the esophagus is so common that the chance of a person's developing this kind of tumor before the age of 65 is about one in six. One of the areas, in Iran, is the Caspian coastal province of Mazenderan; the other, in Soviet Kazakhstan, is centered on a Caspian port city, Guryev. Why are the inhabitants of these areas, nomadic Turkomans in Iran and Kazakh tribesmen in the U.S.S.R., so prone to esophageal cancer?

Since 1966 a team of epidemiologists headed by J. Kmet and E. Mahboubi has been searching for an answer to this question among the Turkomans of Mazenderan. A recent review of their interim findings in *The Lancet* notes that the investigators have been able to rule out genetic factors. Because the Turkomans are strict Moslems, the ingestion of alcohol, which is believed to contribute to the considerable incidence of esophageal cancer in Brittany and Normandy, for example, can also be disregarded. Nor are the Turkomans exposed to any known natural or man-made environmental carcinogens. Kmet, Mahboubi and their colleagues are now focusing on the role that may be played by the mineral deficiencies of the region's soil and the shortcomings of the traditional Turkoman diet. Not only are trace metals absent from the soil in some parts of Mazenderan but also the staple foodstuff, unleavened bread, is rich in a chelating agent that binds metal ions. The Turkoman diet, moreover, is deficient in animal protein and vitamins A, B_2 and C. Iron and B-complex-vitamin deficiencies are believed to make surface cells more sensitive to stimulation by carcinogens, and the Turkomans both smoke and chew tobacco.

Blowing in the Wind

Among the alternative sources of energy that have become more attractive in recent months is the wind. The National Science Foundation and the National Aeronautics and Space Administration have entered into an agreement to build and test a 100-kilowatt windmill-turbine system that would be capable of producing enough electricity to supplement the output of a commercial power plant. On a smaller scale Thomas E. Sweeney of Princeton University is working with the Grumman Aerospace Corporation to develop the Sailwing, a windmill that could produce enough power for a household.

The 100-kilowatt system, consisting of a windmill 125 feet in diameter mounted on a 125-foot tower, will be built at the NASA-Lewis test area in Ohio. The wind speed there averages 10 miles per hour, and speeds of 18 miles per hour or more are recorded each month. The aerogenerator will reach its projected output of 100 kilowatts at a wind speed of 18 miles per hour. The system is expected to generate 180,000 kilowatt-hours of electricity per year in the form of 460-volt, three-phase, 60-cycle alternating current.

The Sailwing bears a resemblance to the sail of a boat. It has a rigid leading edge, tip and root section; the trailing edge is a cable with a Dacron "sail" wrapped around it. Sweeney estimates that the 25-foot experimental Sailwing, together with a generating and storage system, could be made to sell for about $4,000. It is envisioned as a five-kilowatt system; in a place where the average wind velocity is 12 miles per hour the system would produce some 12,500 kilowatt-hours per year, which is about 15 percent more than the average suburban household requires.

Going to the Feelies

The compulsive sybarites of Aldous Huxley's *Brave New World* were entertained by the "Feelies," which added a third sensory channel to the "talkies," themselves a recent innovation when the novel was published in 1932. In Huxley's Utopia one had only to grasp a metal globe to be tactually a party to sexual intercourse on a bearskin rug, "every hair of which . . . could be separately and distinctly felt."

A first tentative step toward the technology of the Feelies has been made by Michael Noll of the Bell Laboratories. For his doctoral dissertation at the Polytechnic Institute of New York, Noll designed and built a "tactile simulation device" that enables one to feel the contours of objects that exist only as mathematical concepts.

Noll's device consists of a large box with a stalk protruding through the top. The box is connected to a computer, and the stalk is fitted with a knob about the size of a billiard ball, which the user holds. When the machine is off or unprogrammed, the knob moves freely in all directions anywhere within a one-foot cube of space. The computer is programmed by entering an equation or a series of equations that describe the object to be simulated; once the program is established, the motion of the knob is limited by the boundaries of the object. When the knob touches the imaginary form, it slides along its surface or bounces off it, as it would if the object were real. By moving the knob throughout its range of excursion one can explore the entire surface of the mathematically constructed form.

Inside the machine are three electric motors and three potentiometers; each controls the motion and senses the position of the knob on one of three axes. The computer continuously calculates the distance of the knob on each axis from the center of the empty cube. When the knob "touches" the edge of the object being simulated, the computer applies power to the motors in such a way that the appropriate resistance or rebound is perceived.

When the machine is operated with the eyes closed, Noll says, the simulated three-dimensional object can be visualized. He suggests that one possible application of the device, which would take advantage of this capability, would be as an aid to the blind.

The present machine simulates tactile perception of the gross features of objects, of what would be felt by moving the hand and arm. It may be possible, Noll believes, to build a more sophisticated device that would also simulate what is felt by the palm and fingertips. Such a machine would provide a sense of texture as well as contour and would more closely approach the aim of the Feelies: to make illusions tangible.

LEON SVIRSKY

Leon Svirsky, who had been managing editor of SCIENTIFIC AMERICAN from 1950 to 1958, died on March 9. His age was 69.

Svirsky was graduated from Yale College, worked as a reporter on the legendary New York *World*, wrote about education and medicine for *Time* and joined SCIENTIFIC AMERICAN at its reorganization in 1948. His main concern at this magazine was the art of storytelling. It was his conviction that any scholar describing his work had it in him to tell a story as gripping as a detective novel. If a particular scholar told a story that was less than gripping, Svirsky felt it was his sworn duty to help him tell it. He was a fierce advocate of a good plot, a good lead and plain English.

Svirsky left SCIENTIFIC AMERICAN in 1958 to work as an editor in book publishing. Later he returned to the magazine, working as a full-time editor from his home in Florida. Five years ago he began to show the symptoms of amyotrophic lateral sclerosis, a progressive and ultimately fatal wasting of the muscles. He continued to work up to the time of his death; he would ask his wife Ruth to lift his hands from the bed up to the typewriter.

THE TOP MILLIMETER OF THE OCEAN

Subtle events that take place in a thin film of liquid covering 70 percent of the earth's surface are decisive for the well-being of terrestrial life

by Ferren MacIntyre

Reduced to the scale of a tabletop globe, the oceans have the thickness of a sheet of onionskin paper. From a planetary point of view the ocean's single most important aspect is neither its depth nor its volume but its surface: the largest and most homogeneous environment on the planet. Through the ocean's 360 million square kilometers of surface pass 70 percent of the solar energy that the earth absorbs, most of its supply of fresh water, a large fraction of the annual production of carbon dioxide and oxygen, a huge tonnage of particulate matter and unmeasured volumes of man-made pollutants. Yet only in the past few years has a significant effort been made to study the ocean surface in detail or to understand how its properties modify the transport of matter or energy across it.

Traditionally oceanographers have defined "surface" as any sample they could catch in a bucket from shipboard. This view is changing as we learn more about the detailed structure of the surface. If one uses a logarithmic scale to plot a cross section of the ocean from the dimensions of a molecule of water lying on the surface to a maximum depth of some 10 kilometers, the top millimeter corresponds exactly to the top half of the ocean [*see illustration on opposite page*]. Thus it is convenient to regard the entire top half of the (logarithmic) ocean as the "surface." The use of a logarithmic depth scale is not simply a literary device; rather it provides a realistic and helpful way of thinking about the complexity of the ocean surface. It makes room for the multiple interacting events that crowd near the surface and helps to clarify their very different depth scales. In terms of events and processes the top millimeter of the ocean offers as rich a field of study as the lower "half" of the ocean.

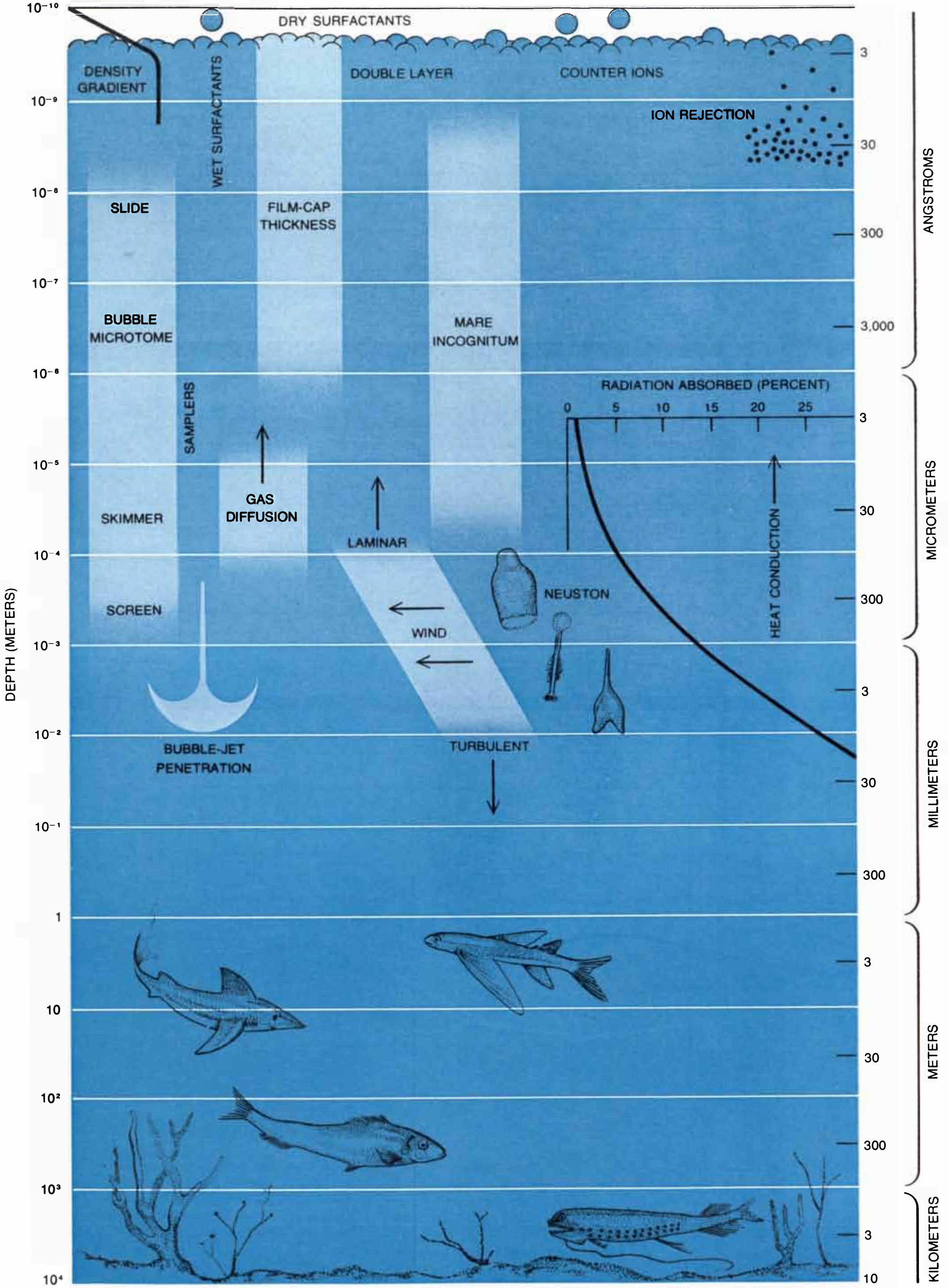
The top millimeter of the ocean has come to be called the microlayer. As an example of the complexity of events in this layer, let us look at the transfer of carbon dioxide across it. The subject is of considerable importance because the burning of fossil fuel will release enough carbon dioxide over the next 20 years to more than double the amount of carbon dioxide in the atmosphere from 320 parts per million today to at least 650 parts per million. No one can predict with confidence how rapidly the excess gas will be absorbed by the ocean and what effect the remnant increase in atmospheric carbon dioxide will have on the world's climate.

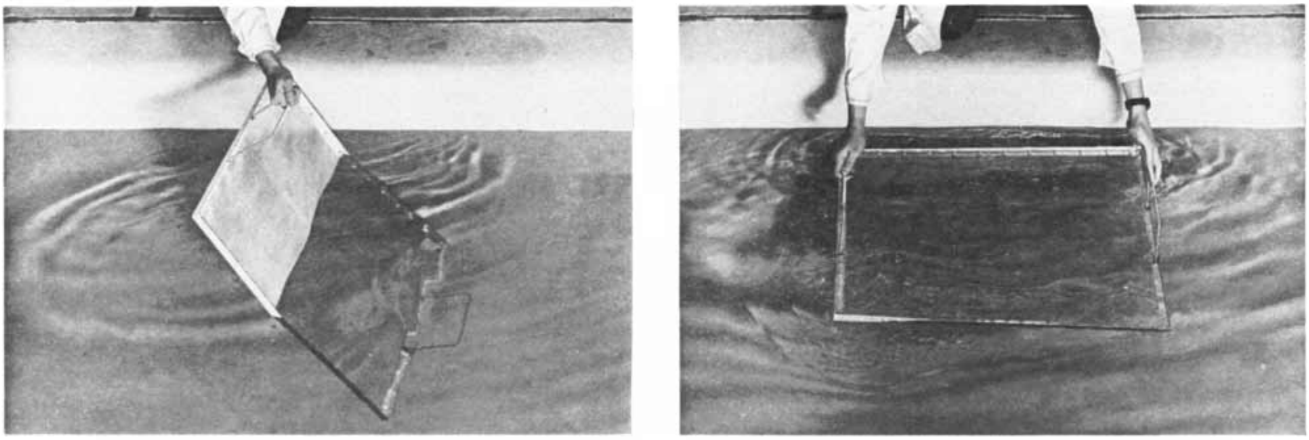
A few years ago reported values of the rate of carbon dioxide transfer across the ocean surface varied by a factor of 330 (if both laboratory and field data were included). The uncertainty reflected our ignorance of the important variables in the process, but gradually things are being sorted out. For example, the rate of transfer across a still surface can be increased fourfold by even the smallest ripples, which have the effect of thinning the laminar boundary layer through which gas must pass by molecular diffusion. Large waves can provide another twofold increase. Paul F. Twitchell of the Office of Naval Research has recently shown that neuston, the tiny plants and animals that live in the microlayer, can increase evaporation (and presumably gas transfer as well) by a factor of three by stirring the laminar layer with their flagella. Of course, the neuston further complicate matters by consuming carbon dioxide as it is passing through the microlayer. These three identified factors still leave a fourteenfold range in transfer rates to be accounted for. One suspects that organic material on the surface (a factor of two?), whitecaps (a factor of three?) and turbulence in both air and sea (another factor of three?) are the most important remaining variables.

We know from estimates of the amount of carbon dioxide released into the atmosphere by the burning of fossil fuels over the past century that the gas disappears from the atmosphere faster than can be accounted for by the known escape routes. The wind-stirred layer of the ocean, the top 100 meters, is more or less in equilibrium with the atmosphere at all times; it can neither absorb more carbon dioxide nor pass it on to the deeper water because there is very little vertical mixing between layers in the ocean. The deep water, representing 90 percent of the ocean volume, has a residence time on the order of 1,000 years. It communicates directly with the atmosphere only in the polar regions, so that it absorbs carbon dioxide very slowly [see "The Circulation of the Abyss," by Henry Stommel; SCIENTIFIC AMERICAN, July, 1958].

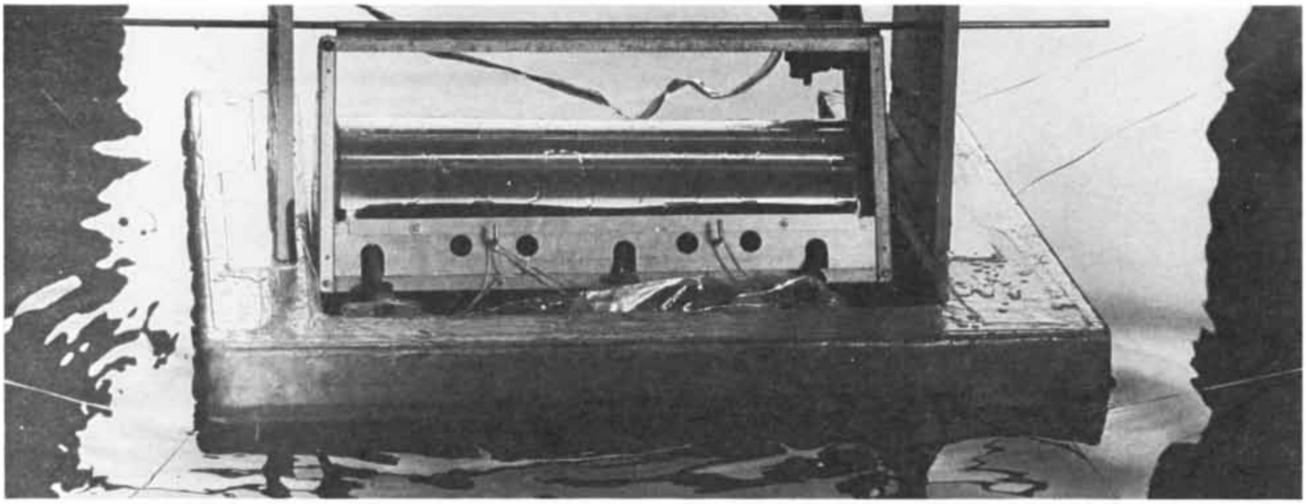
A route into the deep water that may be important, but has not been included

CROSS SECTION OF OCEAN is plotted logarithmically on the opposite page, demonstrating that the top millimeter can be regarded as the upper half of the ocean. The scale starts at the top with the diameter of a water molecule: about 10^{-10} meter, or one angstrom. (The long dimension of a water molecule is actually about two angstroms.) The maximum depth of the ocean is about 10 kilometers, or 10^4 meters. The midpoint of the 14 decades falls at 10^{-3} meter, or one millimeter. Surface chemists study the upper 30 angstroms. Hydrodynamicists study the region between 30 micrometers and one centimeter, where the shearing action of the wind creates boundary layers. The region between 30 angstroms and 30 micrometers is a *mare incognitum*, only now being investigated by microlayer oceanographers.

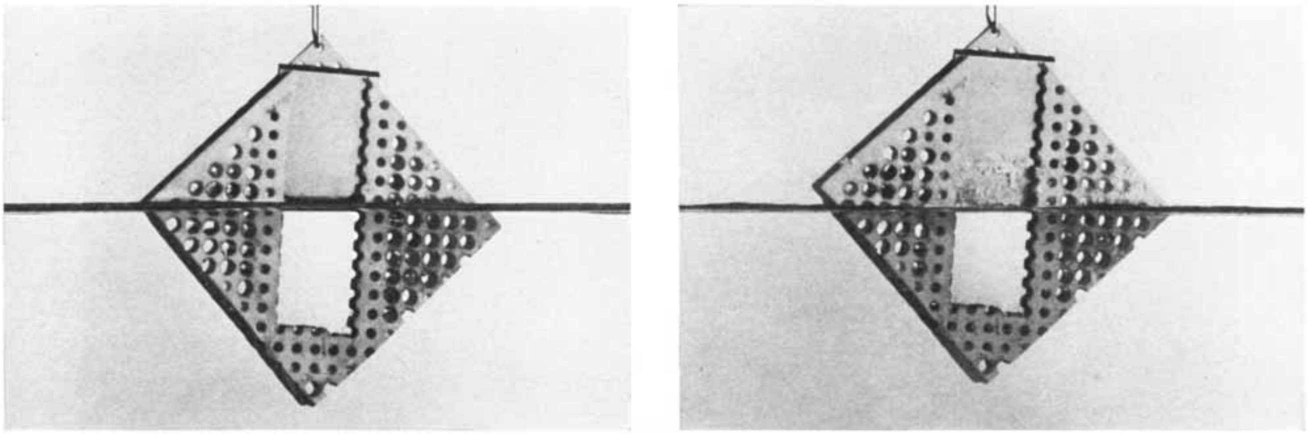

DRY SURFACTANTS
DENSITY GRADIENT
WET SURFACTANTS
DOUBLE LAYER
COUNTER IONS
ION REJECTION
SLIDE
FILM-CAP THICKNESS
BUBBLE MICROTOME
MARE INCOGNITUM
RADIATION ABSORBED (PERCENT)
0
5
10
15
20
25
SAMPLERS
SKIMMER
GAS DIFFUSION
LAMINAR
NEUSTON
HEAT CONDUCTION
SCREEN
WIND
BUBBLE-JET PENETRATION
TURBULENT
DEPTH (METERS)
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
10
10^2
10^3
10^4
ANGSTROMS
3
30
300
3,000
MICROMETERS
3
30
300
MILLIMETERS
3
30
300
METERS
3
30
300
KILOMETERS
3
10



COMPOSITION OF OCEAN SURFACE has proved difficult to establish. The simplest device for sampling the top 300 micrometers of the ocean is a piece of stainless-steel mesh, which can be slipped into the water at an angle and withdrawn parallel to the surface, as is shown here. The device is effective in picking up certain kinds of oily film but leaves many typical surface molecules behind.



MICROLAYER "SKIMMER," devised by George R. Harvey at the Scripps Institution of Oceanography, is a drum coated with a hydrophilic substance. Rolled along the surface, it picks up a continuous film of water 60 to 100 micrometers thick. Both the skimmer and the steel mesh efficiently collect microscopic surface organisms known as the neuston in addition to various "surfactant" materials.



COHERENT LAYER OF SURFACE MOLECULES can be removed by a germanium slide resembling a microscope slide. The device, conceived by Robert E. Baier, adapts a method long used by chemists interested in surface films. At the left, the slide has been immersed. At the right, talc that had been floating on the surface adheres to the withdrawn slide. Baier's contribution was to make the slide of germanium and to bevel the ends so that the surface film adhering to the slide can be examined *in situ*. Since germanium is transparent to infrared radiation, an infrared beam directed into one beveled edge is reflected internally many times before it emerges from the opposite edge. The molecules adhering to the surfaces of the slide selectively absorb characteristic infrared wavelengths so that the emerging beam provides an absorption spectrogram of the adhering substances. All the photographs on this page were made in Baier's laboratory at Calspan Corporation, which was formerly Cornell Aeronautical Laboratory, in Buffalo, N.Y.

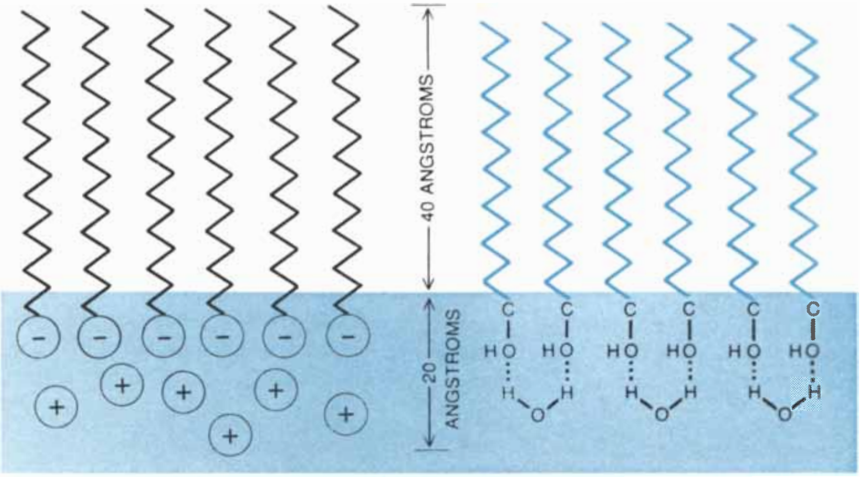
in calculations, is the atmospheric carbon dioxide taken up by the neuston: carbon compounds of fecal pellets and dead organisms sink into the deep water. Once incorporated into a particle large enough to sink, the carbon need fall only 100 meters to short-circuit the normal storage time of carbon dioxide in the mixed layer.

Things That Coat the Surface

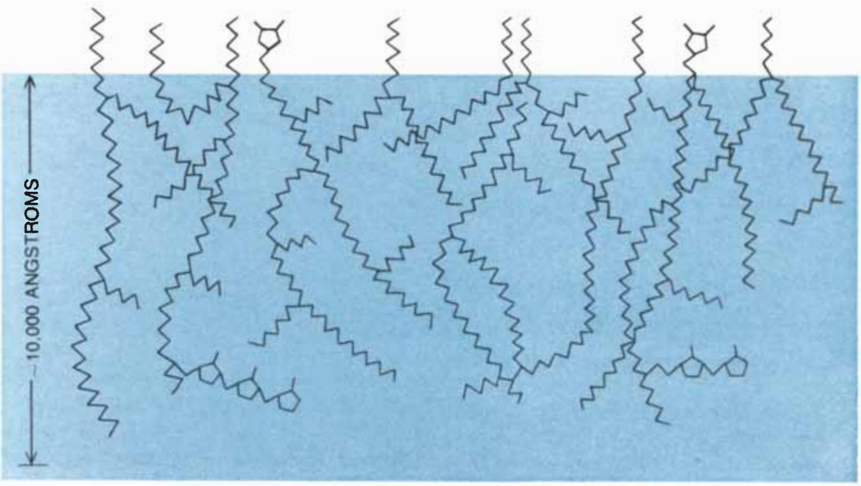
The transfer rates of gas, water vapor and momentum are known to be decreased by minute amounts of surfactants, or surface-active substances. These are long-chain molecules that ubiquitously coat the surface of liquids. The precise nature of the surfactants on the ocean, however, remains an open question. Early workers reported the presence of lipids: molecules that commonly contain 16 or 18 atoms in a hydrophobic tail sticking out of the water and an ionized or otherwise hydrophilic head sticking into the water. These substances, ordinary soap being a typical example, are known as "dry" surfactants because most of their structure sticks out of the water. Robert Morris of the National Institute of Oceanography in Britain reports finding on the ocean such lipid surfactants as hydrocarbons, sterols, esters, glycerides and phospholipids.

On the other hand, Robert E. Baier of the Calspan Corporation (formerly the Cornell Aeronautical Laboratory) finds that lipids are rare on the surface of the ocean except in regions of man-made pollution, whereas what we might call "wet" surfactants are common. They are long-chain proteinaceous substances that are essentially hydrophilic but that stick to the undersurface of the water by virtue of the fact that they have a few hydrophobic side chains [*see illustrations on this page*]. It should be possible to determine the relative amounts of dry and wet surfactants on the ocean surface and whether or not the proportion is changing by developing improved sampling techniques.

It is clearly not easy to invent a method that will efficiently pick up all the substances lying within, say, a micrometer of the ocean surface, or roughly the top half of the top half of the ocean. The earliest microlayer sampler, developed by John M. Sieburth and William Garrett, is a stainless-steel window screen that is immersed perpendicularly to the surface and withdrawn parallel to it. The screen gathers up virtually all surface molecules with a chain length of 16 carbon atoms or more, but its efficiency for collecting shorter molecules may be low.



OILY OR FATTY SURFACE MOLECULES are described as "dry" surfactants because most of the molecule, consisting of a hydrocarbon chain, sticks up above the surface. A typical example is calcium stearate (*left*), which consists of an insoluble tail 17 carbon atoms long that projects upward and a hydrophilic head, a COO^- group, that projects below the surface. Hexadecanol, a 16-carbon alcohol (*right*), forms a liquid monolayer film that is held to the surface by hydrogen bonds linking the terminal OH group with water molecules.



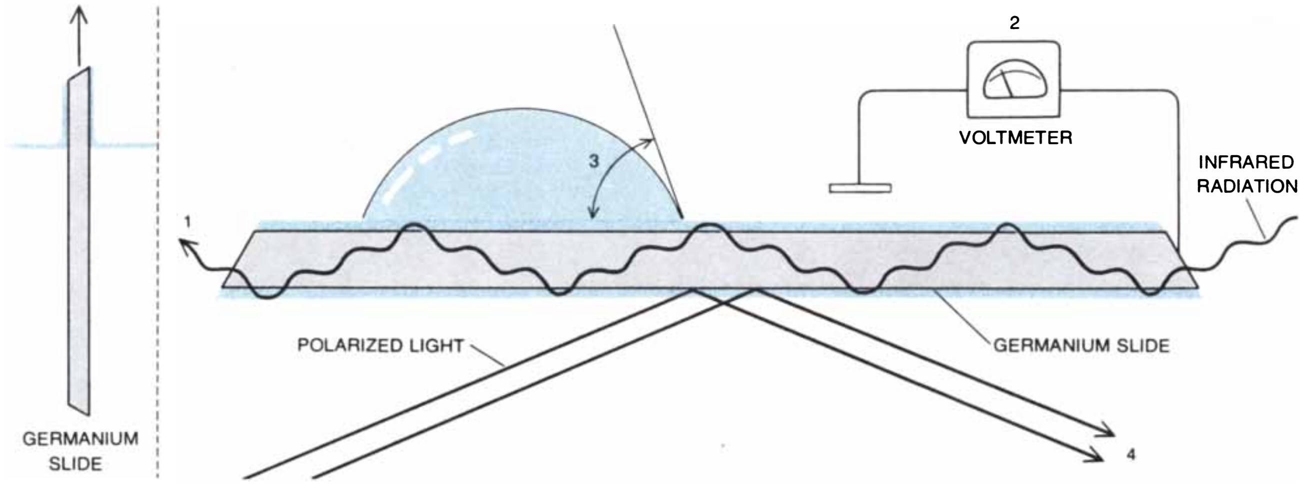
PROTEINACEOUS SURFACE MOLECULES form a broader and more complex class of substances that the author describes as wet surfactants because most of their structure is submerged. Typically proteins and glycoproteins, they are normally much more abundant on the ocean surface than the dry surfactants. Wet surfactants are attached to the surface by the occasional hydrophobic groups that seek to escape from their aqueous environment.

It also collects all the water and living matter lying within about 300 micrometers of the surface. The "skimmer," invented by George R. Harvey at the Scripps Institution of Oceanography, is a hydrophilic drum made of ceramic that rotates slowly on a horizontal axis, with its bottom just dipping into the water. The skimmer picks up a continuous film of water between 60 and 100 micrometers thick, which is scraped off by a blade resembling a windshield wiper.

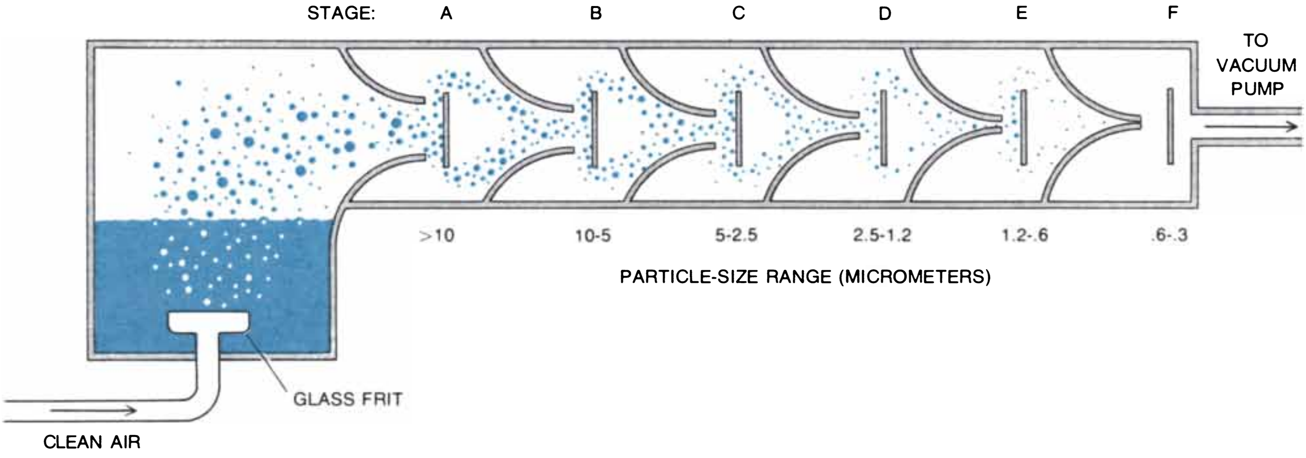
Both of these traditional microlayer samplers collect neuston with high efficiency. Harvey's interest, in fact, lay with this specialized biota; samples taken with the skimmer are sometimes bright green with algae when "deep" water (10 centimeters below the surface) is perfectly clear. In order to extract dry surfactants from a skimmer sample one customarily shakes the sample with chloroform, which dissolves lipids. Unfortunately the chloroform also leaches lipids from the microscopic organisms that are present, confusing the results.

Baier has recently devised an elegant surface sampler that collects no neuston. It is an adaptation of the technique devised years ago by Irving Langmuir and Katharine B. Blodgett for collecting monolayer films from water surfaces. Baier's collector is a piece of germanium shaped like a microscope slide whose ends have been beveled [*see upper illustration below*]. When it is withdrawn perpendicularly from an aqueous solution, it picks up a coherent layer of any surfactants that may be present. The slide is made of germanium because the metal is transparent to infrared radiation. An infrared beam directed into one of the beveled edges of the slide is trapped by total internal reflection until it emerges from the other end. At each reflection the infrared radiation extends about one wavelength beyond the germanium into the adsorbed film of molecules, which absorb particular wavelengths exactly as they would in an infrared spectrophotometer. Thus the emerging beam carries an infrared spectrogram of the functional chemical groups in the adsorbed layer.

Still another microlayer collector coming into use is simply a large funnel, first employed as a sampler by Morris in his study of surfactants. The funnel is filled with surface water and allowed to drain slowly. The neuston float out with the water but the surface film adheres to the sides of the funnel. The film is then rinsed off with a solvent and analyzed by chromatography. Until someone undertakes to calibrate and compare the various samplers there will be no explanation of why the different techniques give such different results. A sampler that has been suggested by several people but not yet tried is simply toilet paper unrolled on the surface of the water. Recognized as a highly efficient way of cleaning a water surface in the laboratory, the toilet-paper method could be used from a dinghy even in moderate waves.

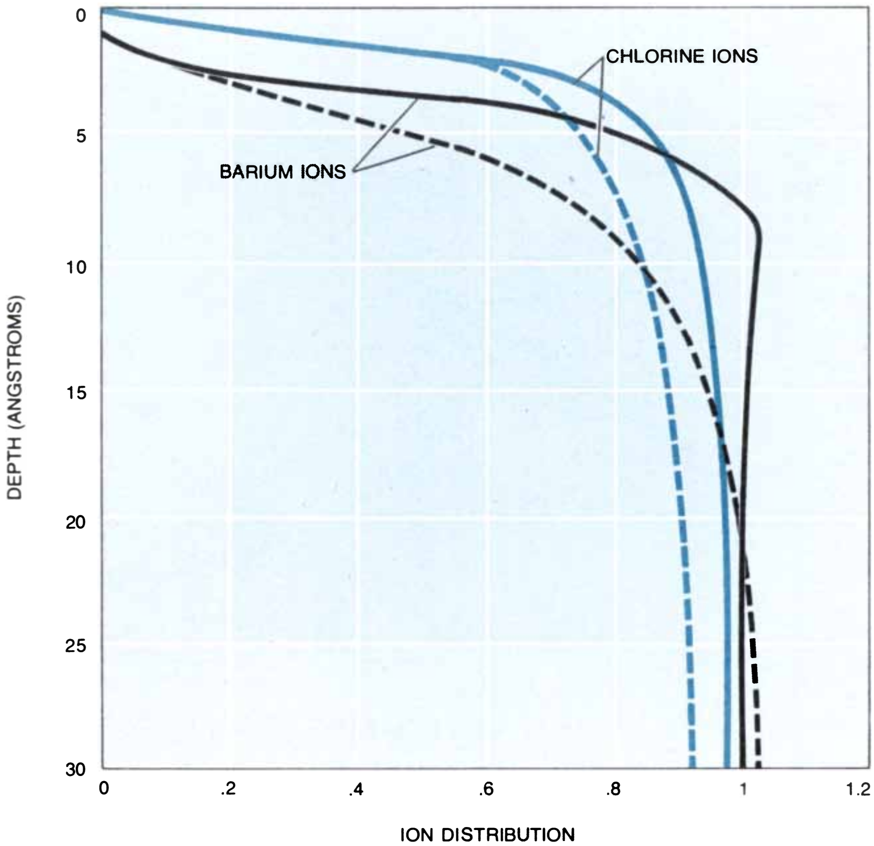


CHARACTER OF SURFACE MOLECULES can be studied in various ways using the sample obtained by Baier's germanium-slide technique. The slide collects a film of both dry and wet surfactants, leaving most organisms behind (*left*). The molecules clinging to the surface selectively absorb infrared radiation, providing clues to molecular structure (*1*). By measuring the surface electrical properties before and after sampling (*2*) one can tell where the molecular dipoles are located and what their density per unit area is. A drop of water placed on the surface assumes a geometry that is dictated by the critical surface tension and hence provides important clues to the free energy and the chemical constitution of the surface. Depending on the surface tension, the tangent angle formed where the drop meets the surface (*3*) can be either greater than or less than 90 degrees. Here it is less, indicating that the film is primarily composed of wet surfactants. The thickness and refractive index of the sample are obtained by reflected polarized light (*4*).



COMPOSITION OF DROPLETS produced when bubbles break the ocean surface is studied in the laboratory with the help of this apparatus. It is found not only that the droplets differ in composition from the bulk sample but also that the composition varies with droplet size. The droplets are sorted as they pass through the stages of an "impactor." Since the air velocity increases at each stage, successively smaller particles are unable to follow the airstream and are collected on impaction plates. In a typical experiment the bulk sample contains two elements whose ratio is carefully determined. The droplets collected on the impactor plates are then analyzed for the same elements. The ratio in the sample divided by the ratio in the original sample yields a value defined as "the fractionation."

The wet surfactants that Baier detects in his samples are not well-known molecules like the dry surfactants, which have been studied by surface chemists for decades. The most intensive work on wet surfactants has been medically oriented and has involved studies of how proteins are denatured (rendered inactive by unfolding) by contact with surfaces. Since the techniques for studying wet surfactants are primitive, almost nothing is known of their thermodynamic, viscous or electrokinetic properties. We have essentially no theoretical understanding of how they affect transport processes through the microlayer except that we can be sure they too retard all forms of transfer. It does appear, however, that the wet surfactants on the ocean, which are known to be glycoproteins and proteoglycans, are reasonably good carriers of phosphate, of various organic molecules, of the scarcer ions of seawater and of heavy metals. Thus they provide a mechanism for trapping and concentrating exotic substances at the ocean surface.



DISTRIBUTION OF IONS AT THE OCEAN SURFACE differs from the average distribution found in the bulk solution. Any substance that increases the surface tension of a liquid would increase the free energy of the surface and therefore must be rejected. The rejection thus operates against the ions of seawater. The theoretical rejection, however, is significant only in the top few angstroms, as these curves for barium and chlorine ions show. The rejection is less for a concentrated solution (*solid curves*) than for one 20 times more dilute (*broken curves*). The concentration of sodium and chlorine ions in seawater is actually five times greater than it is in the stronger of these two solutions. Hence seawater ions would be rejected even less than these minimum values. Thus thermodynamic rejection of ions from the ocean surface cannot explain the distorted composition of seawater particles injected into the atmosphere from the sea. Curves plotted here are based on calculations made by G. M. Bell and P. D. Rangecroft of Chelsea College of Science and Technology in London.

The Puzzle of Rainwater

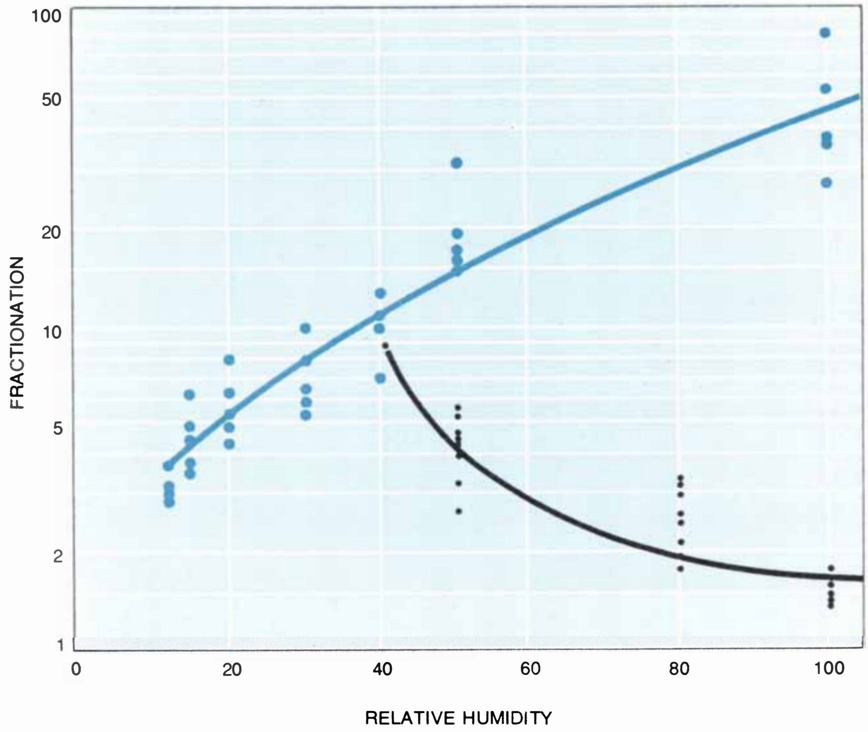
The origins of microlayer oceanography curiously lie in meteorology. After John Dalton had demonstrated in 1822 that sea salt was present in rainwater a number of people suggested that the ocean was the source of other materials found in precipitation. By the early 1940's it was fairly clear that the mixture of marine substances found in rain and snow was not what one would obtain simply by evaporating a typical sample of seawater. In 1959 two meteorologists stated the issue prophetically. "The transport of salts from sea to air," said Carl-Gustaf Rossby, "is by no means a simple mechanical process." Erik Eriksson went on to suggest that "some fractionation process seems to take place at the sea surface, possibly involving material derived from surface films."

At about the same time the Japanese chemist Ken Sugawara reported that when he measured the ratios of iodine and sodium to chlorine in the spray of breaking bubbles, he found values vastly different from the normal ratios in bulk seawater. Sugawara's work was mostly ignored, perhaps because he reported only a single experiment and because the changes in ratios were incredibly large.

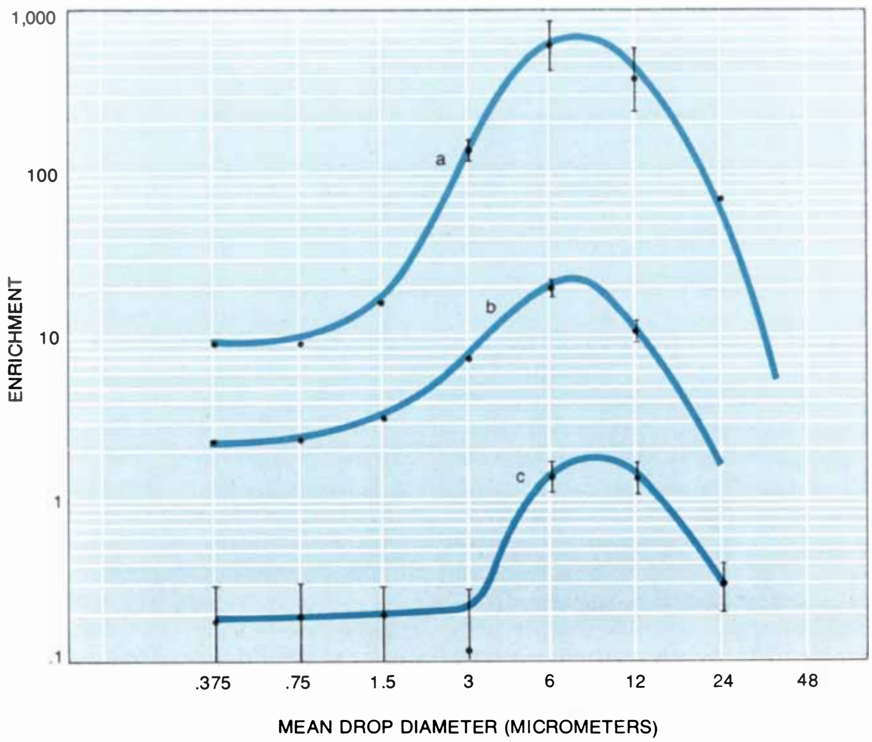
Since then a considerable amount of work on fractionation has shown that if Sugawara had conducted a second experiment, he would have got different numbers, and that his iodine data were surprisingly good. Fractionation experiments, whether they are done in the field or the laboratory, are plagued by nonreproducible results. Fractionation of the principal metallic ions in seawater (sodium, magnesium, calcium and potassium) appears to be slight; if it occurs at all, it alters the seawater ratios by less than 10 percent. It is still impossible to collect a suite of geochemical samples (rain, snow, dry fallout, aerosols and so on) and to estimate the contribution from continental dust with sufficient precision to see a 10 percent variability in the marine component. Gradually, however, from large-scale sampling programs (such as one run by Roger Chesselet in France) our understanding of regional variations in atmospheric constituents is growing, and eventually it may be possible to detect small changes.

Similarly, it is difficult to say anything about fractionation during ejections from the sea surface for substances (such as chlorine, bromine and iodine) that are chemically activated by sunlight or that are catalytically activated in the atmosphere by aerosols. In the laboratory few workers have appreciated how readily fractionation experiments can be confused by surface-active impurities and how extraordinarily difficult it is to prepare a clean surface. Just once have I been able to keep a surface clean for 24 hours. It took a year of preparation, and even then it was mostly good luck, since I never succeeded again. The experience taught me one useful fact: reproducible results are possible only when surface conditions are reproducible. Impurities of a few parts per billion can change the surface properties of water; at the part-per-million level impurities can alter fractionation results so completely as to mask any of the often sought physicochemical mechanisms.

Early searches for the fractionation



LARGE FRACTIONATION VALUES for potassium ions (*black*) and ammonium ions (*color*), far exceeding any possible fractionation based on ion rejection, have been obtained by M. R. Bloch of the Negev Institute for Arid Zone Research. The experiments were conducted using a version of the impactor illustrated at the bottom of page 66. The correlation between humidity and fractionation is very puzzling and is so far unexplained.



PHOSPHATE FRACTIONATION VALUES obtained by the author demonstrate the dramatic influence of trace amounts of surfactant molecules in experiments using the drop-impactor technique. The water used in the experiment that yielded curve *a* was assumed to be carefully distilled. Actually it seems to have contained something that strongly attracted phosphate ions to the surface. (The vertical coordinate, "enrichment," is simply "fractionation" minus 1.) Curve *b* was obtained after several hours of bubbling had removed the phosphate-attracting contaminant. Curve *c* shows how fractionation was depressed when a negatively charged surfactant was added to repel the negative phosphate ions.

mechanisms at the sea surface ignored the unique properties of the top half of the ocean and relied instead on traditional physicochemical thinking. More than 100 years ago Josiah Willard Gibbs had shown that any substance that increases the surface tension of water will tend to be rejected as it diffuses upward from below. Since different salt solutions have different surface tensions, it follows that their ions are differentially rejected, so that the surface of a mixture such as seawater will differ in composition from the bulk. Although the difference is calculable, it extends into the liquid only a small fraction of a micrometer and seems far too small to be of any geochemical importance [*see illustration on preceding page*].

Nevertheless, some experiments designed to test the Gibbs equations reveal a startling disagreement with theory, both in the scale of fractionation and in the direction. For example, whereas theory predicts that a solution containing barium and cesium ions should be slightly depleted in barium ions immediately below the surface, amounting to a fractionation of about .005 percent, experiment shows a barium enrichment amounting to a fractionation of about 7 percent. This remarkable discrepancy is unexplained.

In other experiments M. R. Bloch of the Negev Institute for Arid Zone Research in Israel has found that under nonequilibrium conditions the fractionation of potassium and ammonium ions is strongly influenced by the humidity above the liquid surface [*see top illustration at left*]. Potassium in bubble spray is enriched as much as tenfold and ammonium as much as a hundredfold compared with their concentration in the bulk sample. The extraordinary values for ammonium ions (NH_4^+) are perhaps related to the hydrolytic formation of ammonia (NH_4OH), which, being uncharged, is not rejected from the surface by electrical effects. (It may be relevant that the dielectric constant of surface water has a value of only 9 whereas for water in bulk it is 81. This suggests that one might profitably study fractionation in solvents whose dielectric constant resembles that of surface water.) In any case it remains a mystery, at least to me, how humidity alone can produce the striking effects Bloch reports. There is also a question of whether even an extreme degree of fractionation within the range of electrical forces—say a few hundred angstroms from the surface—would be significantly reflected in the composition of film drops thrown up as an aerosol. Perhaps the flow of heat and water vapor under nonequilibrium conditions are capable of altering the composition of the film cap as it is being formed, but this is a difficult question to approach either theoretically or experimentally.

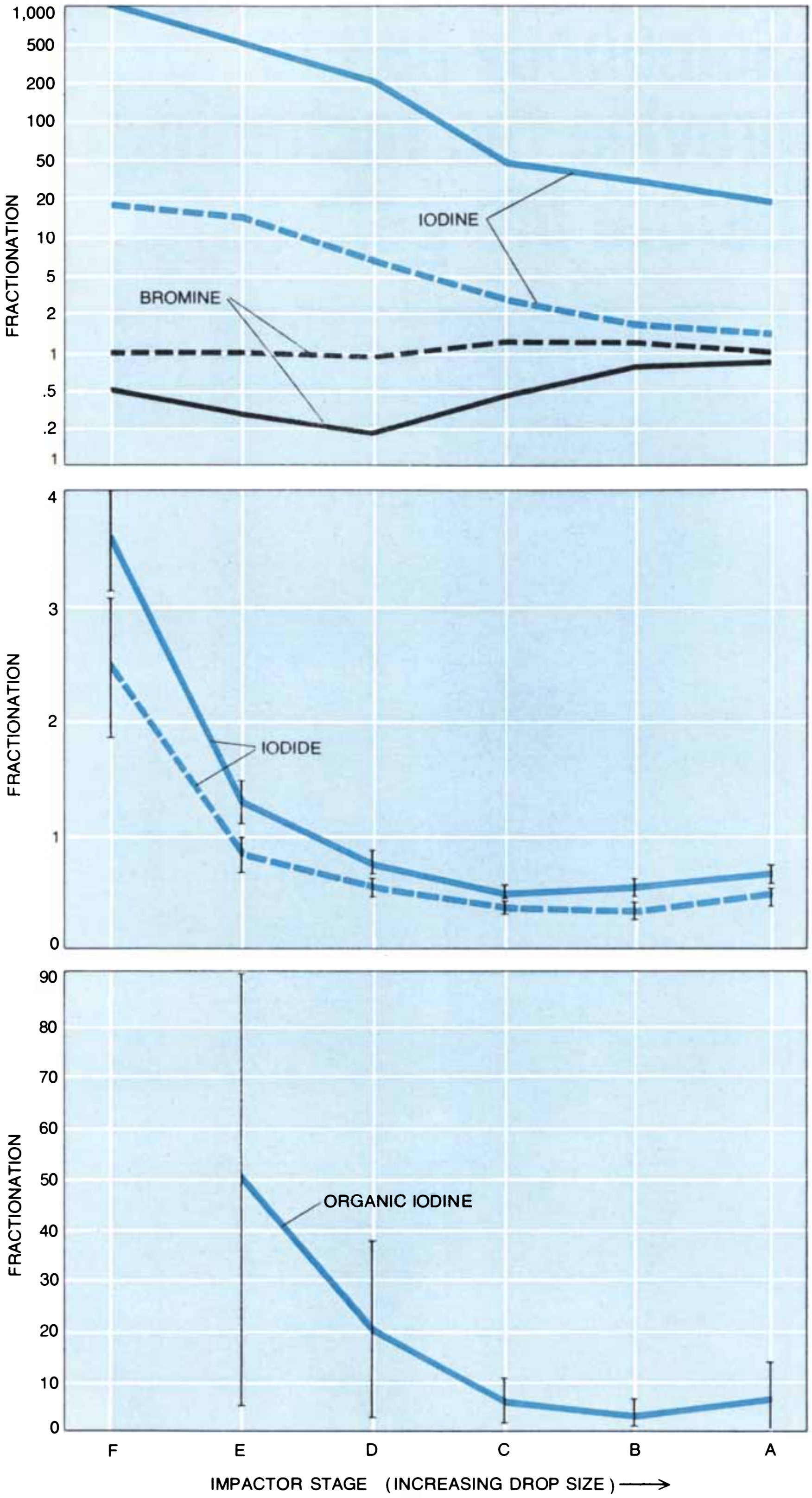
Some of the most spectacular values of fractionation have been obtained in experiments with phosphate and iodine. In my own early work I correlated the fractionation of phosphate with drop size, using the aerosol-impactor technique shown in the lower illustration on page 66, and I found that the enrichment can be as high as 600 times and is rarely less than three [*see bottom illustration on opposite page*]. It was also clear that the degree of enrichment could be influenced by the presence of surfactants.

The fractionation of iodine and other halogens has been intensively studied by Robert A. Duce and his co-workers at the University of Hawaii and at the University of Rhode Island. Duce's colleague F. Y. B. Seto finds that iodine is rejected from large aerosol particles but that enrichment increases rapidly for particles that are smaller than about two micrometers. He also finds that if the particles are exposed to sunlight, a significant fraction of the iodine escapes as a gas [*see illustration at right*]. Seto's results are the only reproducible fractionation work of which I am aware. It is obvious that the large fractionation values observed for iodine cannot be caused by Gibbsian rejection of ions within a few angstroms of the surface.

The Recipe for Rainwater

The clue to the entire mystery of what happens at the surface was provided, in my opinion, by G. A. Dean of the New Zealand Medical Research Council, who was fascinated by the composition of rain. In 1963 he analyzed the rain at Taita, a small town on the New Zealand coast, and published a "recipe" for duplicating its content of the major ions found in seawater. The recipe called for .5 milliliter of seawater (to supply sodium, potassium, magnesium and calcium in their normal ratios), four milligrams of dried plankton and algae (to supply iodine, phosphorus, nitrogen and additional potassium), together with enough distilled water to make one liter. Dean's recipe provided a seminal insight into the relation between macroscopic meteorology and microlayer oceanography, by emphasizing the tremendous amount of biological material that accompanies the small particulate fraction (.05 percent) of the total water that moves from sea to air.

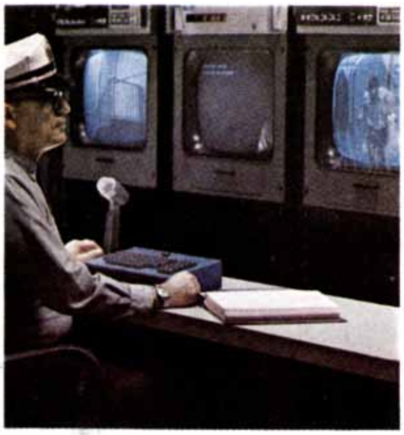
The biological material in Dean's



FRACTIONATION OF OCEANIC IODINE is geochemically important. The top set of curves shows that old oceanic aerosol from the trade winds (*solid lines*) contains more iodine and less bromine than aerosols freshly generated by ocean surf (*broken lines*). Bromine is unfractionated initially and is subsequently lost to the atmosphere. Iodine is not only fractionated initially but also gradually becomes concentrated in smaller drops. The two curves in the middle diagram, based on laboratory studies by F. Y. B. Seto of the University of Hawaii, demonstrate that iodide is rejected from large particles. The broken curve shows the additional effect of exposing droplets to 18 minutes of ultraviolet radiation, which volatilizes part of the iodine. The bottom curve shows that plankton grown with radioactive iodine yield organic material that collects preferentially on the small droplets.

Somebody has to provide the technology before life gets safer

Safer from killer storms.
You're looking at a tropical storm south of the Baja Peninsula from an altitude of 450 miles. The left picture, from a 2-mile resolution infrared sensor, shows temperature of the scene. The right picture, obtained simultaneously with the same resolution, shows reflected visual light. The combined information from both pictures gives the type, height, and vertical contour of the clouds. This Westinghouse multisensor system also includes a low light-level sensor that sees under moonlight conditions. Now, unequaled pictures of the earth give better warning of hurricanes, tornadoes, heavy rains and snows—even fog.



Safer from saboteurs and vandals.

Our integrated security systems are designed to protect key facilities and valuable assets against theft, fire, and malicious destruction. They combine perimeter and internal intrusion detection, with surveillance cameras that "see in the dark." We tie these systems together through a centralized command and control center. We install and service equipment, and train your employees, too.

Safer in the air.

We're building some of the world's most advanced air traffic control equipment. Westinghouse provides display equipment, surveillance radars, instrument landing systems, and beacons for use throughout the Federal Aviation Administration's control network. This technology, coupled with skilled air traffic controllers, makes your flight safer from takeoff to touchdown.

You can be sure if it's Westinghouse. For our brochure, write Marketing Manager, Defense, Westinghouse Electric Corporation, Pittsburgh, Pennsylvania 15222.

Westinghouse helps make it happen

Polaroid's SX-70: The New Photography.

The Polaroid SX-70 Land camera atop a single section of the original electronic layout. Ultraminiaturization has compressed hundreds of electronic components into three small brain centers.

No matter how you've looked upon the act of taking pictures, instant or otherwise, picture-taking will never be the same.

A color photograph, which develops before your eyes outside the camera, is only the *beginning* of the SX-70 experience.

The rebirth of a sense of wonder and an increased awareness of the beauty that surrounds you are the end results of SX-70 picture-taking.

Watch it happen.

Remarkable as the Polaroid SX-70 Land camera is, what is important to remember is not what *it* can do, but what it enables *you* to do.

Touch the red electric shutter button and—whoosh—the picture is automatically ejected from the front of the camera.

There is nothing to time, nothing to peel, nothing to throw away.

Your picture begins developing immediately beneath a transparent, protective plastic cover. You can stack it with other pictures, handle it, put it in your pocket. You can even spill water on it and still not hinder development.

As the image blooms before your eyes, you realize that this will be a color photograph such as you have never seen before.

Even after you have a beautiful picture, it keeps getting better.

Minutes later, you're looking at your finished picture. It is hard, dry, shiny, flat and extremely durable.

The image size of the picture is 3⅛ x 3⅛ inches.

The camera that makes it possible.

The SX-70 Land camera, closed, is 1 x 4 x 7 inches. Wrapped in top-grain leather, it weighs only 24 ounces.

Inside, there exist over 200 transistors, a complex system of fixed and pivoting mirrors and a 12,000 r.p.m. motor, all working together to propel one brilliant picture after another into your life.

Yet all you have to do is frame, focus and touch the red electric button.

The SX-70 is a single lens reflex camera. Your eye sees precisely what the camera lens sees, because you're actually looking through the camera lens.

You can focus on a scene from miles away, down to 10.4 inches, and you can actually record far more detail than you can see without the use of a magnifying lens.

You can shoot up to 10 pictures, one every 1.5 seconds, to cover an action sequence, like a baby's first jaunt across the room. Or you can take automatic time exposures up to 14 seconds long.

The scope of SX-70 picture-taking is as wide as your imagination will allow it to be.

Power from the film pack.

As amazing as the Polaroid SX-70 Land camera is, the story of the other

Touch the button, your film is automatically ejected.

In seconds, you see the faint outline of your picture. Your picture, minutes later, fully developed.

half of the system, the film, is equally magical.

Each 10-exposure film pack is small enough to be put into your pocket, yet each contains a wafer-thin, 6-volt battery that provides fresh power to operate the camera every time you load fresh film.

A unique picture counter, located on the back of the camera, tells you how many pictures you have left. When you insert your film pack, the counter reads "10." After each successive shot, that number decreases. It even prevents the flash from firing after all 10 pictures have been taken.

Flash pictures.

Just as there are 10 exposures in each film pack, there are 10 shots on each GE FlashBar™ array (5 on each side). You can shoot flash pictures from 10 inches to 20 feet or more away. Once again, you can reshoot every 1.5 seconds, to get a full action sequence.

Nothing being left to chance, the SX-70 electronically selects the next flash to be fired. And, as you focus, the lens aperture is set automatically to let in the correct amount of light.

Outdoors, an electric eye reads the light and programs the electronic shutter for the correct exposure.

Copies and enlargements.

Because SX-70 pictures can record far more detail than the unaided eye can see, and because they are reproduced in a whole new spectrum of vibrant, highly stable dyes, beautiful copies and enlargements are easily made.

Every time you load new film, a cover sheet is automatically ejected from the front of the camera, on which you'll find directions on how to obtain copies and enlargements of your favorite SX-70 original pictures.

The new photography.

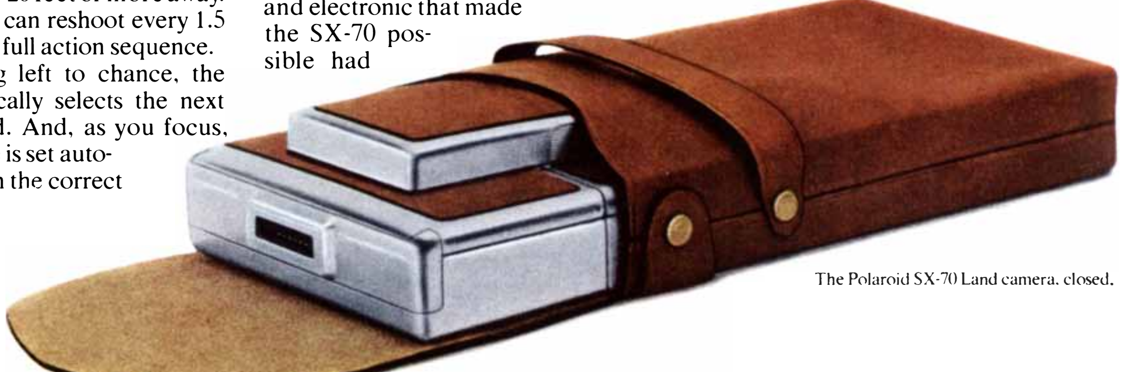
The virtual cascade of revolutions, mechanical, chemical, optical and electronic that made the SX-70 possible had only one purpose: to free you from everything cumbersome and tedious about picture-taking, so it could become at last, the simple creative act it should be.

Just frame, focus and touch the red electric shutter button. Your picture is automatically ejected to time itself and develop into a color photograph of a depth and brilliance unparalleled in amateur photography.

What once might have seemed a family duty, or even just an interesting hobby, can now become a spontaneous and recurring pleasure in your daily life.

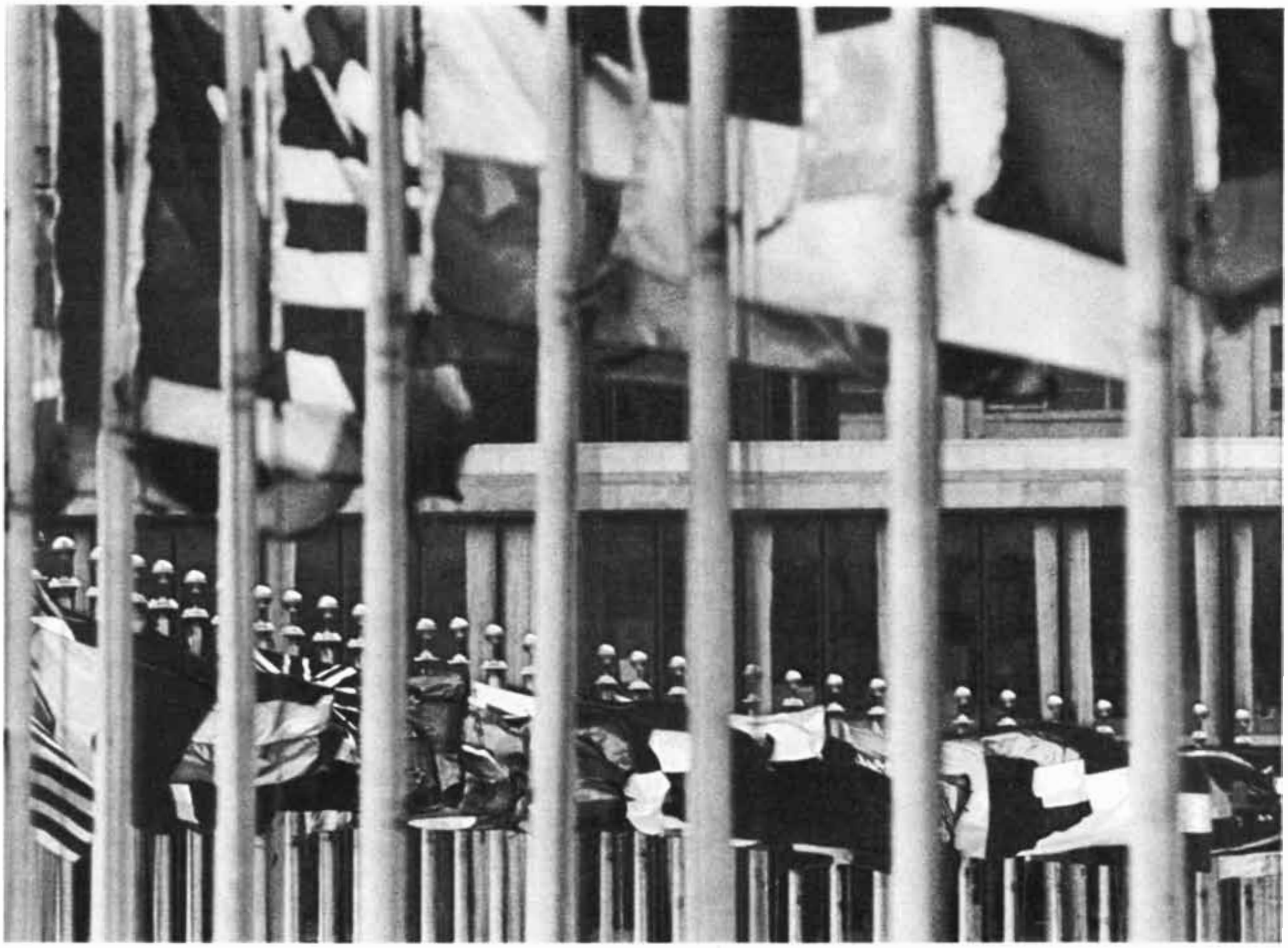
The price.

The suggested list price of the Polaroid SX-70 Land camera is $180. It is now available at Polaroid camera dealers.



The Polaroid SX-70 Land camera, closed.

Polaroid®



If you make anything that uses electrical connectors –anywhere in the world–call AMP.

We can help you put a smile on your bottom line.

If you manufacture computers, automobiles, appliances, telephones or anything else that uses electrical connections—from Tokyo to Toronto to Turin—maybe we can help brighten your profit picture.

We have some 100 plants, warehouses and sales offices throughout the world supplying highly reliable AMP electrical and electronic connectors.

This includes everything from simple, one-piece contacts to sophisticated interconnection systems. Plus a full line of economical application machines for virtually every major industry.

So if you make a product that uses electrical or electronic connections—anywhere in the world—give AMP a call at (717) 564-0100. Or write for our new brochure: "We can help you put a smile on your bottom line."

AMP Incorporated, Harrisburg, Pa. 17105.

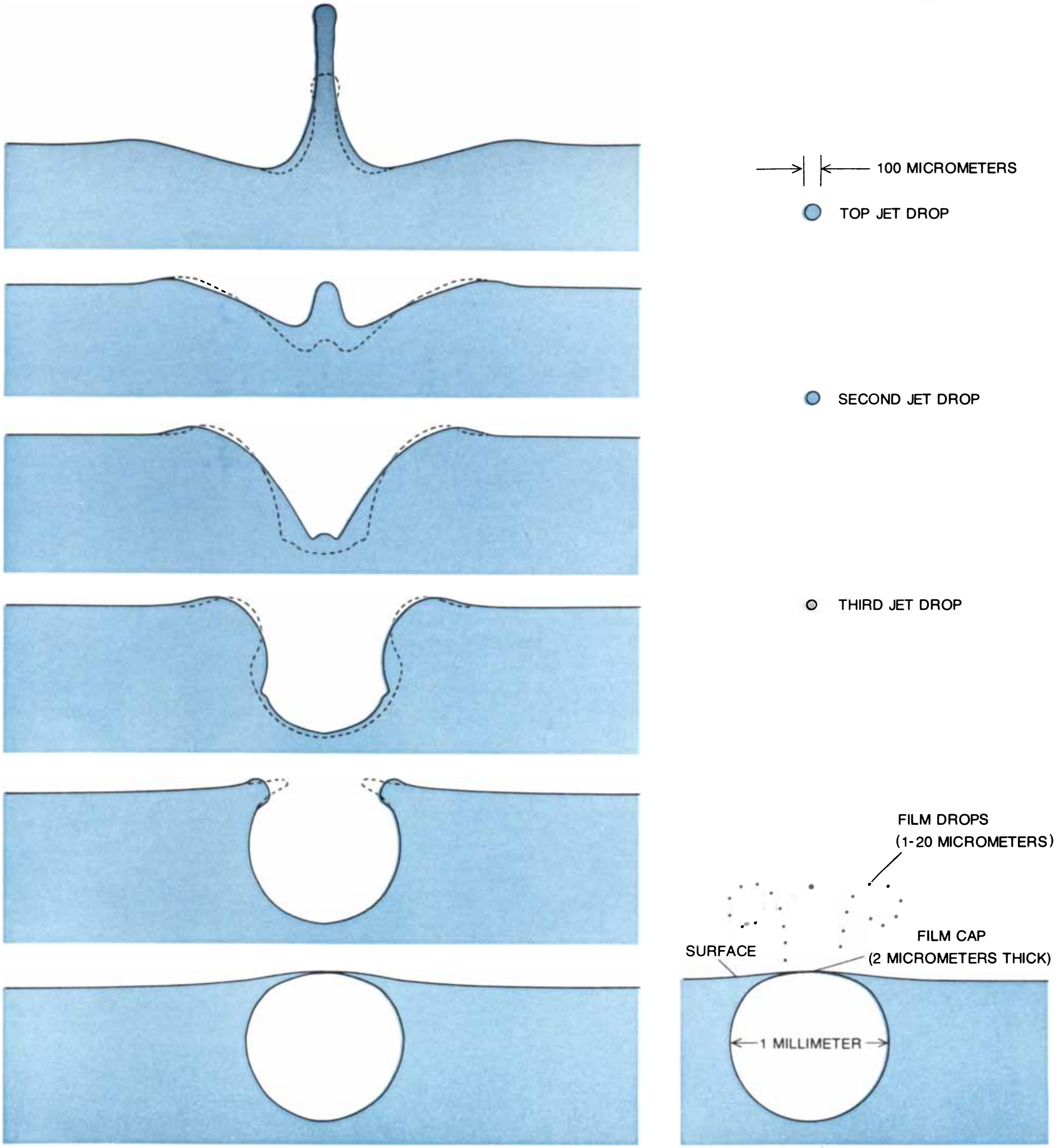
Facts you should know about AMP.

- Over $400 million annually in sales.
- Over 10% spent on Research, Development and Engineering.
- Over 1,000 field service and sales engineers.
- Some 100 manufacturing plants, warehousing facilities and sales offices in USA and worldwide: Barcelona, Brussels, Buenos Aires, Frankfurt, London, Lucerne, Mexico City, Montreal, Paris, Puerto Rico, San Paulo, s'Hertogenbosch (Holland), Sydney, Stockholm, Tokyo, Toronto, Turin, Vienna.

AMP
INCORPORATED

recipe, approaching the concentration found in a weak bouillabaisse, represents 2,000 times the amount of organics normally present in seawater. How does that much biological material find its way into the atmosphere? Duncan C. Blanchard, who was then working at the Massachusetts Institute of Technology, showed that drops from breaking bubbles could carry off an organic monolayer from a liquid surface, and I estimated that the thickness of sea surface so skimmed off could hardly be more than a tenth of a micrometer thick. At the time no one was able to see how so thin a slice could be ejected into the air.

Since then, by piecing together half a dozen arguments (no one of which is entirely convincing by itself), I am persuaded that a breaking bubble is a surface microtome of great finesse. It is



BREAKING BUBBLES, according to the author's hypothesis, provide the chief mechanism for injecting into the atmosphere a peculiar selection of the substances present in the top micrometer of the ocean. The six drawings at the left show how a collapsing bubble projects a high-velocity jet of liquid into the air. The sequence is based on high-speed photographs of a 1.7-millimeter bubble made by Duncan C. Blanchard, A. H. Woodcock and others at the Woods Hole Oceanographic Institution. The drawings depict the collapse at intervals of 1/3,000th second; profiles represented by broken lines show an intermediate stage 1/6,000th second earlier. The acceleration of surface material can reach 1,000 times gravity (1,000 *g*) for a 1.7-millimeter bubble and as much as a million *g* for a 10-micrometer bubble. The drawing at right shows schematically the variety of drops produced by a bursting bubble. Roughly a fourth of the available energy is carried off by the top drop of the jet, which then frequently evaporates into an airborne particle of salt.

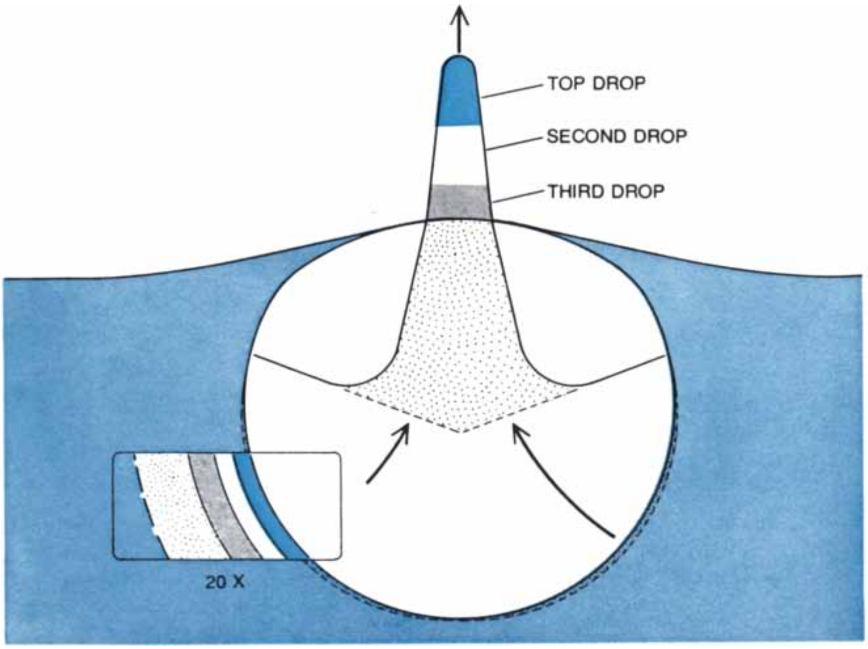
capable of removing a sample of ocean surface as thin as a tenth of a micrometer, or about 1,000 times thinner than the best mechanical microtome yet devised for removing a slice of liquid.

High-speed photographs of breaking bubbles made at the Woods Hole Oceanographic Institution indicate that when a 1.7-millimeter bubble bursts, it can accelerate surface material to a value 1,000 times greater than the acceleration produced by the earth's gravity [*see illustration on preceding page*]. For bubbles of 10 micrometers the acceleration of surface material can reach a million *g*. Intuition is not much help in thinking about the consequences of such events. An analysis suggests that the average thickness of ocean surface ejected in the atmosphere by a bubble microtome is roughly .05 percent of the bubble diameter. A bubble one millimeter in diameter will remove a slice of ocean surface about .5 micrometer thick [*see illustrations on these two pages*].

The fastest-moving drop produced by the collapsing of a one-millimeter bubble is ejected upward at about 10 meters per second, carrying about a fourth of the energy released by the collapse. The drop reaches a height of between 10 and 15 centimeters and frequently evaporates into an airborne particle containing 30 nanograms of salt and perhaps .3 nanogram of dried plankton and algae. The organic material on the ejected drops is compressed, so that if the bubble interior is covered by a monolayer of organic molecules, the drop will be wrapped in perhaps 10 layers. The process delivers economically significant amounts of plant nutrients to coastal areas and, along with pollen and droplets of condensed volatiles from terrestrial plants, contributes enough food to high-altitude snow to support invertebrate life in the "aeolian" zone above the vegetation line. The organic material tossed up from the sea also makes the windshield of your car greasy if you park near the ocean and in times of "red tide" can include enough toxin to cause sore throats and eye irritation among shore dwellers.

The Ubiquity of Bubbles

Some 3 or 4 percent of the sea surface is covered with bubbles at any moment. Assuming that 10^{18} break per second, between one and 10 billion tons of salt are ejected into the atmosphere annually. Each drop also carries about 200 charges of positive electricity. The charge results from the peculiarities of flow at the moment the drop separates from the ascending jet. Air friction decelerates the surface of the jet by values that can reach 500,000 *g* while the interior is still moving rapidly upward; the resulting shear across the electrical double layer injects a few excess positive ions into the drop. Over the world ocean this amounts to a steady upward current of some 160 amperes. The excess of positive ions may account in part for the general feeling of well-being one experiences at the seashore.



SOURCE OF DROP MATERIAL in a breaking bubble is the thin wall of the bubble just before collapse. The twentyfold enlargement of the bubble wall is color-keyed to show the origin of the top, second and third drops that will ultimately break away from the rising jet.

Workers at the Woods Hole Oceanographic Institution have found that when an inert gas is bubbled through seawater, "dissolved inorganic" phosphate is converted into "particulate organic" phosphate. Brine shrimp will grow slowly when given the phosphate-containing particles as their sole food. Very few marine organisms can extract dissolved material from seawater, and even they do so with great difficulty, but since every particle will be eaten by something, the particle-forming process is important in making phosphate (which is often a limiting nutrient) available to the plankton community.

The mechanism by which bubbles induce the formation of particles was the subject of much controversy. This seems to have been resolved by the finding that very small particles are necessary as nuclei. Apparently if the nuclei are adsorbed onto a bubble, the surface compression that occurs when the bubble breaks results in a permanent transfer of sea surfactant onto the nucleus, whereas in the absence of the nucleus the surfactant simply clumps and redissolves almost immediately. Similar particle growth (and dissolution) can be observed by half-filling a horizontal glass tube with soap solution. When the tube is tipped upright, thereby reducing the surface to a small fraction of its original area, surface compression creates a lump of soap out of what had previously been a monolayer.

A related phenomenon is observed when a gas free of carbon dioxide is bubbled through seawater. Since the gas carries off the dissolved carbon dioxide, equilibrium then demands that calcium carbonate precipitate out of solution. The seawater becomes opalescent with fine particles, a phenomenon known as "whiting" when it occurs over the Bahama Banks (through a different mechanism). The particles precipitated in this fashion are unique among carbonates in that they do not dissolve in acids. The reason seems to be that they form at the bubble surface, where the carbon dioxide concentration is lowest and the carbonate supersaturation is the highest, so that when the bubble breaks, the particles are coated with a tenacious film of organic material that protects them from acid attack. Instant Saran Wrap!

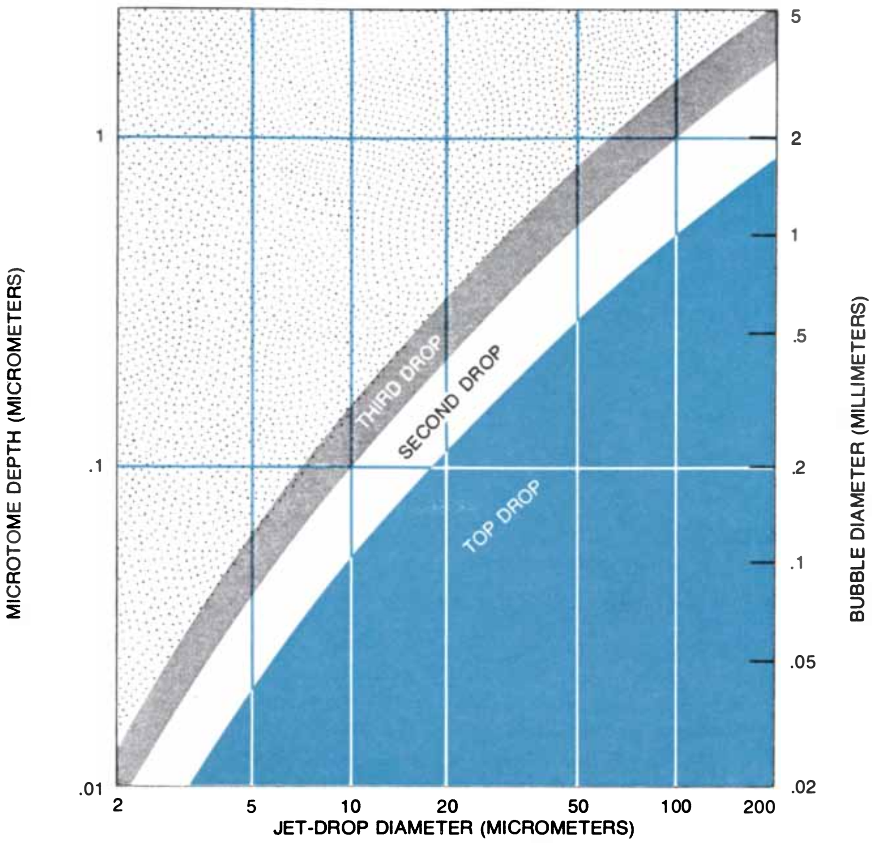
Pollution and Sea Farming

Not only is the oceanic microlayer exposed to all the pollutants in the atmosphere but also it absorbs the brunt of oil spills, including oil released from blowouts on the ocean floor. Some 85 percent of all the organic material now afloat on the Baltic, the Mediterranean and North seas consists of petroleum. The microlayer avidly concentrates such heavy metals as lead, mercury, copper, chromium and zinc. It also retains such long-lasting chlorinated hydrocarbons as DDT and PCB (polychlorinated biphenyls), which reach the sea in precipitation and surface runoff. These substances enter the food chain, where they are progressively concentrated.

Perhaps the most bizarre example of oceanic dispersal of a pollutant is the one deduced to explain the appearance of DDT in Antarctica long before it could be carried there by wind or water currents. The sequence starts with the spraying of DDT on Northern Hemisphere crops. A portion enters the atmosphere by codistilling with the water that is transpired through the leaves of plants. The DDT reaches the North Atlantic in precipitation, where it is trapped by the microlayer. It promptly enters the food chain and is concentrated as it passes from smaller organisms to larger ones. The agent for transmitting the DDT across the Equator is Wilson's petrel (perhaps the most abundant bird in the world), which feeds in the North Atlantic but breeds in the Antarctic. So transported, the DDT is picked up by the food chain in the southern ocean. When it finally reaches the tissues of the crabeater seal, it appears for the first time in concentrations high enough to be detected.

I shall close by mentioning two ways in which increased understanding of microlayer oceanography may be of great practical significance. The first relates to the potential control of hurricanes. It is suspected that both water vapor and salt nuclei from the sea play a crucial role in the spawning and growth of hurricanes. Since both substances must cross the microlayer, it is conceivable that making small and inexpensive changes in the sea surface ahead of an incipient hurricane might reduce the storm's severity. With this in mind one investigator is studying the evaporation rate from the microlayer, having noticed that there are anomalies in the surface properties of seawater at about 30 degrees Celsius, the same ocean temperature known to favor the genesis of hurricanes.

A second way in which the microlayer might be manipulated for man's benefit would be as a source of protein-rich food. One reason agriculture has so greatly outstripped mariculture is the extensive use of fertilizer on land. Fertilizing the sea has never seemed profitable because both fertilizer and crop can escape so easily. (The sustained yield of the semienclosed North Sea fishery depends entirely on the nutrients poured in by the Seine, the Thames, the Rhine and other open sewers—a fortuitous and nonoptimized bit of mariculture.)



BUBBLE "MICROTOME" can produce a thinner slice of the ocean surface than any mechanical device known. The average thickness of the sample ejected into the atmosphere by a breaking bubble (*scale at left*) is about .05 percent of the bubble diameter. Thus the collapse of a bubble one millimeter in diameter will produce a sample in the top jet representing roughly the top .5 micrometer (5,000 angstroms). Subsequent jet drops, all about 100 micrometers in diameter, originate in onionlike shells successively deeper in the solution.

One can imagine an optimized microlayer farming operation designed along the following lines. Microlayer "weeds," or unwanted organisms, are first removed with a "cultivator," which is simply a Harvey drum skimmer. Surfactant fertilizer, growth hormone and "seed" (desirable varieties of neuston) are next applied. The fertilizer could be designed as a recyclable nutrient carrier that could not escape the microlayer, rather than as a totally consumable molecule. Since each step of the food chain loses 90 percent of the energy contained in the previous step, it is advantageous to harvest the phytoneuston (the plant species) before they are eaten by the zooneuston (the animal ones). Optimum operation of the farm would have most of the nutrients converted into biomass by the time the prevailing wind had moved the microlayer down to the harvesting skimmer. Even so, grazing predators might have to be fenced out. The farm should be about as productive as a Kansas cornfield or a Caribbean sugar plantation.

Obviously many problems need to be solved before a microlayer farm is feasible. Present efforts at collecting oil spills might be directly applicable to harvesting. The operation of a farm would have to be largely automatic, utilizing power available on site from the sun or the wind. The vagaries of weather are no greater at sea than on land. And it might prove feasible to protect the farming apparatus by submerging it 10 meters or so during storms.

Perhaps the biggest advantage of mariculture over agriculture is the ease with which microscopic organisms can be manipulated genetically. Strains of *Chlorella* have already been bred that are 80 percent oil, whereas cottonseed, our richest present source, is only 20 percent oil. Similar results might be achieved with protein content. The food-processing industry should be clever enough to convert neuston into a product both palatable and nutritious.

GLYCOPROTEINS

Many proteins have sugars linked to them as side chains. The function of these saccharide groups is not known in all cases, but they may radically alter the properties of the molecule

by Nathan Sharon

Proteins are linear polymers; their long strands of amino acids may be twisted or pleated or folded, but they are never branched; one strand is never connected to the middle of another to form a side chain. Side chains of another kind, however, are quite commonly linked to proteins. They consist of sugars, and the substances they form in combination with proteins are called glycoproteins, from the Greek *glykys*, sweet.

Until recently sugar found in association with most proteins was assumed to be an impurity, and measures were taken to separate it from the protein. Only in the slimy proteins, or mucins, found in external secretions such as saliva, was sugar an acknowledged constituent; the covalent bonding of carbohydrates to proteins in these substances was recognized more than 100 years ago. It has required improved methods of isolation, purification and identification, however, to demonstrate that mixed polymers of sugars and amino acids are of wide distribution in nature and participate in many biological processes. In the past 10 years it has become apparent that the majority of proteins are in fact glycoproteins.

Among the proteins known to have side chains of sugars are most of the proteins of the blood plasma, including the substances that determine human blood types, and a large number of enzymes and hormones. Collagen, a structural protein important in the formation of cartilage and other connective tissue, is also a glycoprotein. Interferon, an antiviral agent produced by the body, is another; that it contains bonded sugars was demonstrated only last year. Glycoproteins have also been detected in the brain and in nerve endings and may participate in some unknown way in memory and in the transmission of nerve impulses.

Another group of glycoproteins, found in the blood serum of several species of Antarctic fish, act as antifreeze agents. They are remarkably effective, lowering the freezing point of aqueous solutions to a greater degree than equal weights of sodium chloride.

Glycoproteins are found in plants as well as animals, as my colleagues and I at the Weizmann Institute of Science in Israel demonstrated 10 years ago in studies of soybean agglutinin. Many other plant agglutinins, or lectins, which are characterized by their ability to make cells clump together, have since been shown to contain carbohydrates. More recently glycoproteins have also been discovered in some bacteria and viruses.

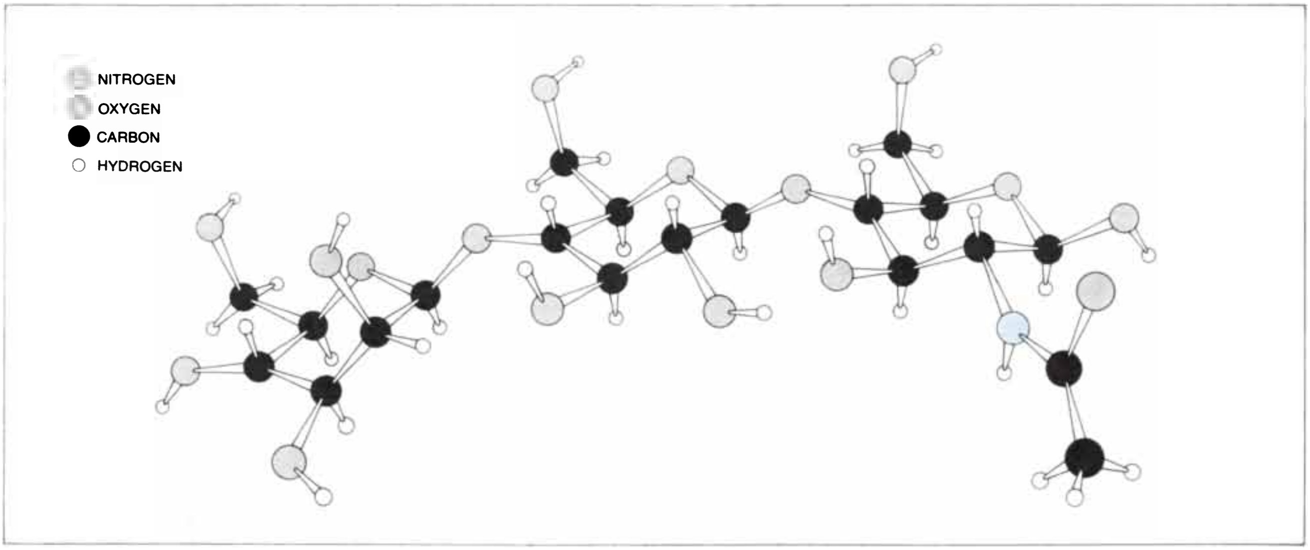
The most intriguing group of glycoproteins now being investigated are those found on the surfaces of cells. Although in mass they are only minor constituents of the membranes of animal cells, they are important because they are responsible, at least in part, for communicating the individuality of the cell. It has been said that all cells come with a sugar coating. Most of this "coating" is in the form of glycoproteins, which serve as antigenic determinants, as virus receptors and as markers of cellular identity. This last function may prove to be particularly significant, since it has recently been determined that the carbohydrates on the cell surface are altered when cells become malignant.

The basic units of proteins are amino acids; those of sugars are monosaccharides. In proteins the carboxyl group (COOH) of one amino acid is connected, with the loss of a water molecule, to the amine group (NH_2) of the next; this linkage is called a peptide bond, and proteins are therefore polypeptides. There are 20 common amino acids; the sequence in which they are combined entirely determines the properties of the protein.

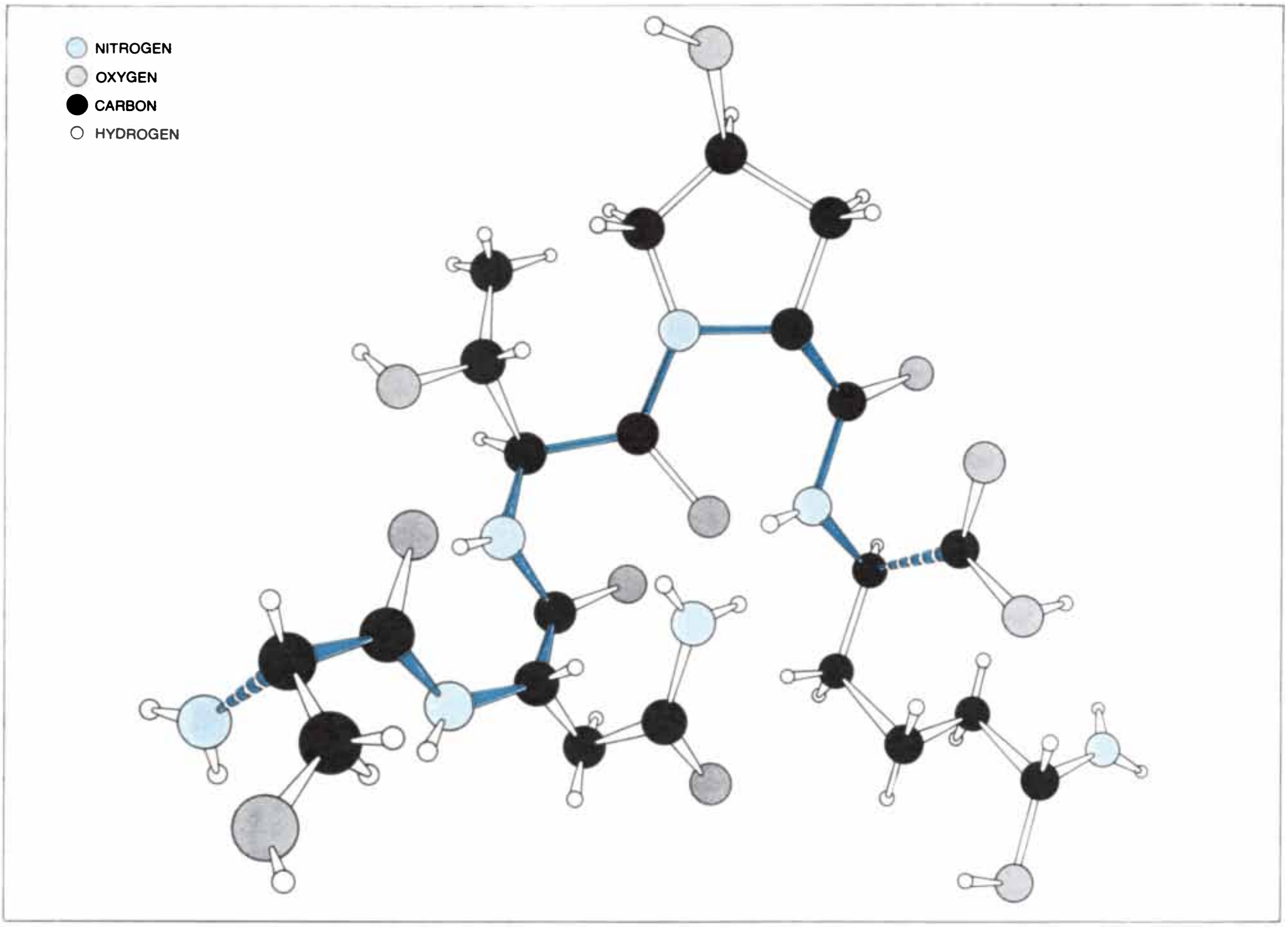
The formation of sugar polymers is potentially much more complex. More than 100 monosaccharides are known in nature, and they can combine with one another in many ways, rather than in only one way as in proteins. Only nine of these monosaccharides have been detected in glycoproteins, and they are usually combined in oligosaccharides of no more than 15 sugar units; the number of possible combinations is nevertheless still very large.

Glucose, a ubiquitous metabolite and one of the most thoroughly studied of all organic compounds, is found in glycoproteins, but not frequently; it appears to be confined mainly to collagens. Galactose and mannose, which are somewhat more common in glycoproteins, are stereoisomers of glucose, that is, their chemical formula is the same, but some of its elements are oriented differently. In fucose, which is related to galactose, a hydrogen atom replaces one hydroxyl group; in acetylglucosamine and acetylgalactosamine an aminoacetyl group takes the place of one hydroxyl [*see illustration on page 80*].

All these sugars are hexoses; the essential portion of their structure contains six carbon atoms. Two other glycoprotein monosaccharides, arabinose and xylose, are pentoses; each has five carbon atoms. Of considerably different structure is neuraminic acid (in certain forms called sialic acid), a sugar bearing both aminoacetyl and carboxyl groups. Sialic acid was recognized as a common constituent of glycoproteins in the 1950's as a result of the work of Alfred Gottschalk, who was then at the Walter and Eliza Hall Institute of Medical Research in Melbourne. Because of its acidity this



SUGAR COMPONENT OF GLYCOPROTEINS consists of monosaccharides, or individual sugar units, linked together to form oligosaccharides, or chains of a few units each. All the monosaccharides found in glycoproteins have in common a central structure called the pyranose ring, a group of five carbon atoms with the first and fifth connected by an oxygen atom. The three-dimensional form of the pyranose ring is often compared to the shape of a chair; it is bent to form three planes, like a chaise longue. Attached to this central structure are hydroxyl groups and other groups that to a large extent determine the characteristics of the sugar. The oligosaccharide above consists of (*left to right*) the monosaccharide units mannose, glucose and acetylglucosamine.



PROTEINS ARE ASSEMBLED from amino acids, the carboxyl group (COOH) of one being attached to the amine group (NH_2) of the next by the extraction of a water molecule. The hypothetical segment of a protein above is formed from five amino acids. They are (*left to right*) serine, asparagine, threonine, hydroxyproline and hydroxylysine, the five amino acids that carbohydrates are known to link with in glycoproteins. The main chain of the protein is indicated by solid color bonds; if this chain is traced from left to right, the bonds between amino acids (the peptide bonds) are those preceding the nitrogen atoms. Most proteins have 100 to 1,000 amino acid units; all terminate in an amine group (*dotted bond at left*) and a carboxyl group (*dotted bond at right*).

sugar endows the polymers it forms with distinctive chemical and physical properties. The discovery that it is present on cell surfaces and that it acts as the receptor for the influenza virus was a fundamental stimulus to the study of glycoproteins.

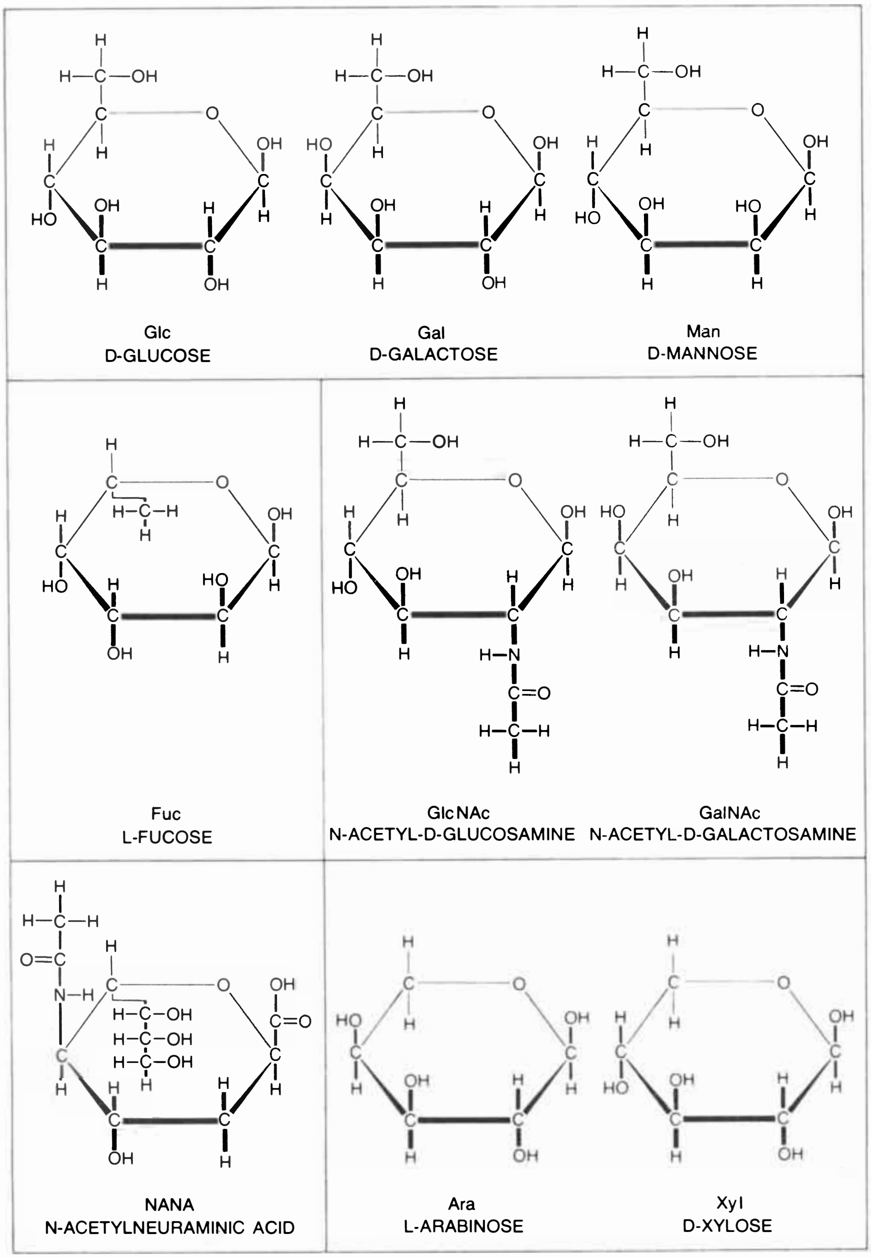
The known glycoproteins range in size from a molecular weight of about 15,000 to a molecular weight of more than a million, and they may contain from 1 percent to more than 85 percent carbohydrate. Ovalbumin, for example, the major protein of hen egg white (molecular weight 45,000), and the enzyme ribonuclease B (molecular weight 14,700) each have only one carbohydrate side chain per molecule. At the other extreme are proteins such as sheep submaxillary mucin (molecular weight about a million), each molecule of which incorporates some 800 sugar units. These units are arranged in disaccharide side chains and are densely distributed, with a disaccharide for every six or seven amino acids.

Even in glycoproteins incorporating only a few sugar residues an enormous number of saccharide structures are possible. Two glucose molecules, for example, can form 11 isomeric disaccharides, since they can be bonded at any of five positions and at four of those positions they can be linked in either of two ways and at the fifth in any of three ways. Two identical amino acids, on the other hand, can form only one dipeptide. Three different sugar molecules can be arranged in 1,056 trisaccharides, whereas three different amino acids can form only six tripeptides.

As the number of sugar residues increases, the number of possible structures soon becomes overwhelming. One reason for this complexity is that oligosaccharides, unlike proteins, can form branched chains of molecules, and at each point where the chain divides, the side groups can be linked to the main strand in a number of ways. John Clamp of the University of Bristol has calculated the number of possible isomers in a saccharide side chain somewhat like that of α_1 acid glycoprotein, a component of human blood plasma. The molecule he considered consisted of three residues each of mannose, acetylglucosamine, galactose and neuraminic acid, all linked to another acetylglucosamine residue that provided the connection with the protein. If only the composition of this oligosaccharide is known, and not the structure, there are about 10^{24} possible isomers. Even if the sequence of sugars is known, there remain some 6×10^6 possible bonding arrangements.

The complex structure and the large number of possible structures of oligosaccharides complicate the task of analyzing them. Furthermore, the methods available for the investigation of carbohydrate polymers are not as refined as those that have been developed for the elucidation of protein and nucleic acid structure.

The information sought in the analysis of glycoproteins includes the composition and sequence of amino acids in the peptide chain, the composition of the sugar constituents, the number of saccharide chains, their structure, the nature of the carbohydrate-peptide linkage and the position of the carbohydrate side chains on the peptide sequence. The number of glycoproteins for which all this information is available is very small,



SUGARS FOUND IN GLYCOPROTEINS are depicted schematically, with the pyranose ring compressed into a single plane. Glucose, galactose and mannose are stereoisomers; they differ only in the orientation of the groups attached to one of their carbon atoms. Fucose is related to these sugars but has a hydrogen atom substituted for a hydroxyl group; acetylglucosamine and acetylgalactosamine have aminoacetyl groups at another position. Acetylneuraminic acid bears both aminoacetyl and carboxyl groups; arabinose and xylose lack the sixth carbon of the glucose-derived sugars. *D* and *L* stand for *dextro* and *levo* rotatory and refer to the orientation of the groups attached to four atoms in the pyranose ring.

but the composition and the partial structure of many have been determined. From these results certain general patterns of glycoprotein construction can be inferred.

It is known, for example, that fucose and sialic acid always occupy peripheral positions, farthest from the polypeptide chain. Acetylglucosamine and galactose, on the other hand, are often found nearest the protein and often form part of the carbohydrate-peptide linkage. The nature of this linkage itself has also been extensively explored. Of the 20 common amino acids only five are known to form linkages with carbohydrates. They are asparagine, serine, threonine, hydroxylysine and hydroxyproline; the last two are amino acids to which a hydroxyl group has been appended after the synthesis of the protein [*see illustration on this page*].

The first linkage to be identified was that between acetylglucosamine and asparagine. The bonded sugar and amino acid were isolated from ovalbumin in 1963 after a long, sustained effort by Albert Neuberger and Robin Marshall of St. Mary's Hospital in London; this linkage was identified in the same year in other laboratories in the U.S., Japan and the U.S.S.R.

The first step in analyzing a glycoprotein is to isolate and purify it. Every effort is made to preserve it in its undenatured state so that its biological activity is not impaired. The presence of carbohydrates in the molecule offers an additional "handle" with which the glycoprotein can be carried through the processes of fractionation and by which the purity and homogeneity of the final product can be assessed.

The purified glycoprotein is digested exhaustively with proteolytic enzymes, which cleave peptide bonds and thus break the protein into fragments. The smaller fragments, consisting of one amino acid or a few amino acids with a saccharide side chain attached, are then isolated. From these glycopeptides the sugars can be removed sequentially by treatment with various glycosidases, enzymes that cleave the glycosidic bond and thereby remove specific sugars one at a time from the end of an oligosaccharide or polysaccharide [*see illustration on page 84*].

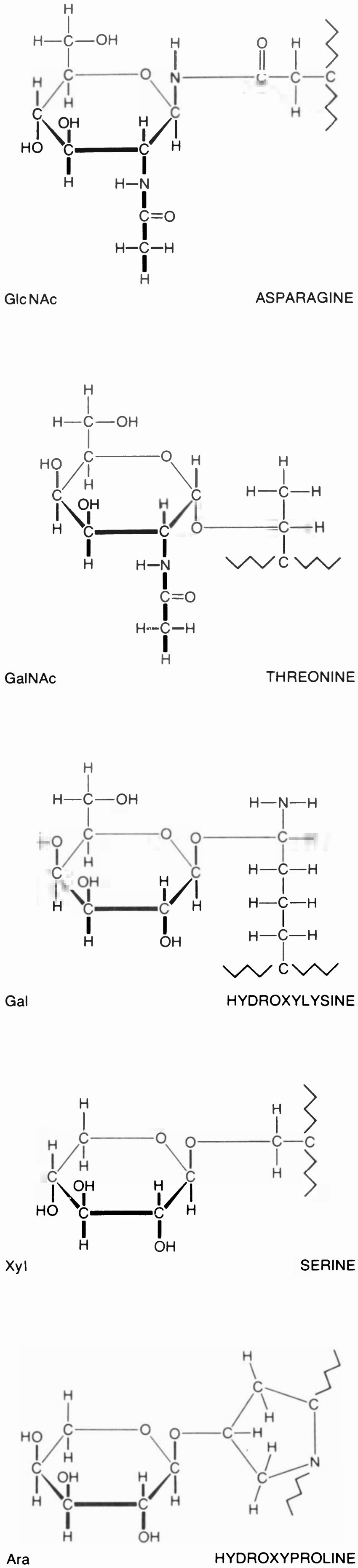
Ideally the end product of such a serial degradation is a compound composed of a single amino acid linked to a single sugar, for example asparagine and acetylglucosamine. This was in fact the product we obtained from soybean agglutinin; we were therefore able to conclude that the carbohydrate-peptide linkage in this plant glycoprotein is the same as that found in many animal substances.

Enzymatic degradation of this type can yield information about the structure of the oligosaccharide side chains as well as about the carbohydrate-peptide linkage. For example, if a glycopeptide is initially incubated with a mannosidase and mannose is liberated, one can conclude that mannose residues are present on the periphery of the molecule. Subsequent incubation with other glycosidases will reveal the identity of the next unit in the chain, and so on. From the evidence provided by this method, and from other techniques of analysis, the structure of several glycoprotein saccharides has been established [*see illustration on next two pages*].

In proteins amino acids are always present in integral ratios. In many glycoproteins the ratio of one sugar residue to another is not an integral one. For example, in ovalbumin the ratio of mannose to acetylglucosamine is about 5.5 : 4. The difference between peptides and saccharides represented by this observation is fundamental to the nature of both kinds of compounds.

Within an organism all the molecules of any given protein are identical; consequently the ratios of amino acids calculated from any sample of the protein will be the same as those present in a single molecule. Carbohydrates, however, often vary in structure and composition from molecule to molecule. The formula for an oligosaccharide established by analysis is therefore an average or typical structure and any given molecule may differ from it. Even in relatively simple glycoproteins having only a single carbohydrate side chain, the saccharides are heterogeneous. In fact, nonuniform side chains have been detected in ovalbumin derived from a single egg of a pure-bred hen.

The heterogeneity of the carbohydrates of glycoproteins is a direct consequence of the mechanism of their syn-



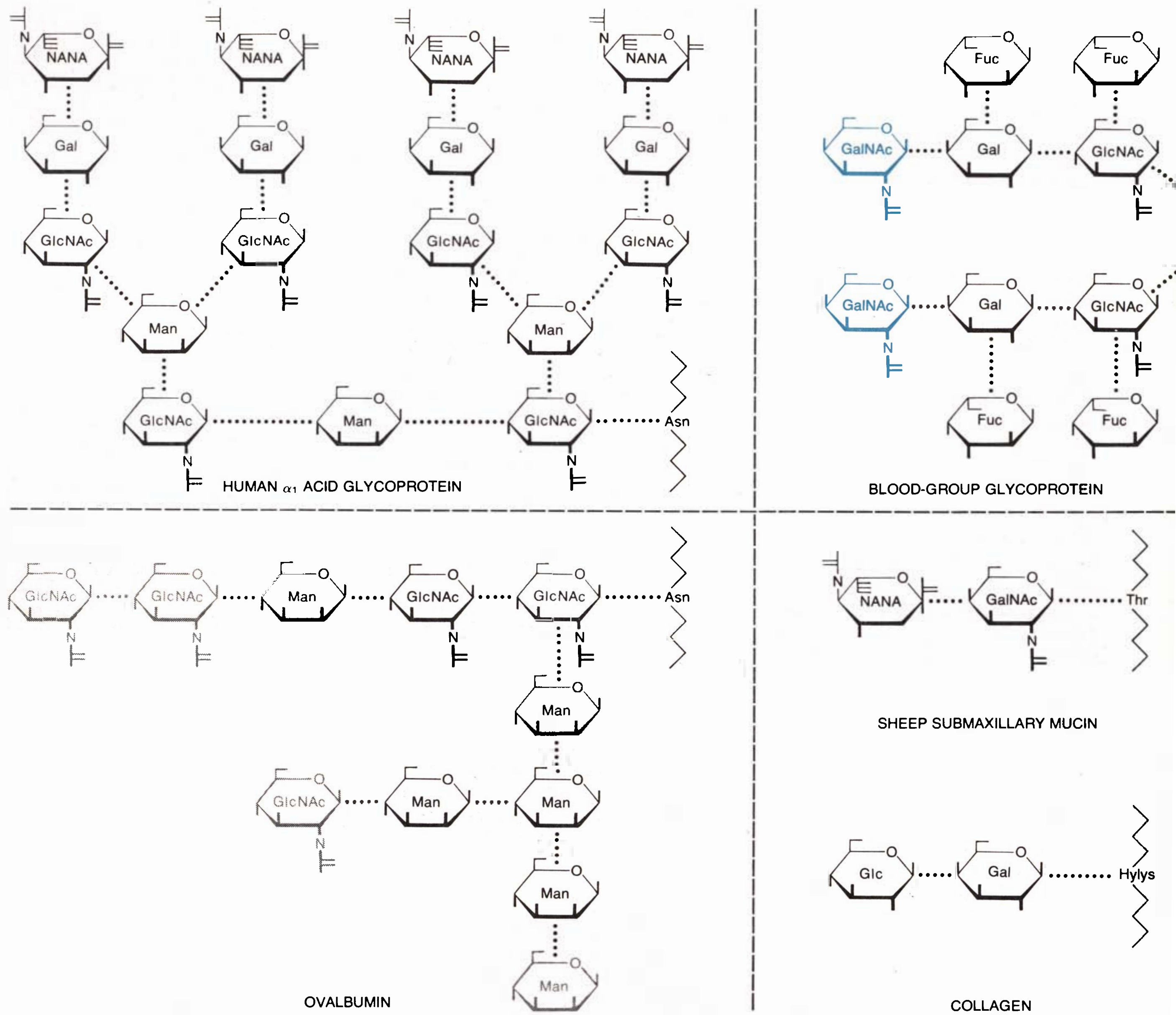
BONDING of sugars to proteins is known to involve only five monosaccharides and five amino acids. All the linkages are through glycosidic bonds, which are formed by the extraction of a water molecule. The acetylglucosamine-asparagine linkage is through a nitrogen atom; all the others are through oxygen atoms. Acetylgalactosamine can bind to either threonine, as shown, or to serine.

thesis. This process begins with the formation of the polypeptide part of the molecule; like all other proteins it is assembled at the ribosome on a "template" represented by DNA and RNA. As a result its structure can change only through the mutation of the genetic material of the cell. The provenance of proteins in the genetic code is also responsible for their invariant linearity. DNA and RNA are themselves linear polymers, and the code they embody dictates merely sequences of amino acids; it has no provisions for governing the branching of the polypeptide chain.

The carbohydrate component of a glycoprotein is not a product of the ribosome; it is added to the completed polypeptide at another site in the cell, by another process. The details of this process have become apparent during the past 10 years as a result of the work of Saul Roseman of Johns Hopkins University, Robert G. Spiro of the Harvard Medical School and many others.

Carbohydrate side chains are synthesized at sites on membranes; the process is mediated by enzymes specific to each saccharide rather than by a genetically determined template. As a result much more variation is possible in the product than in the formation of the polypeptide component of glycoproteins. Where RNA specifies that a particular amino acid *must* be attached, the enzymatic mechanism responsible for the formation of oligosaccharides specifies only that a particular monosaccharide *can* be attached [*see illustration on page 85*].

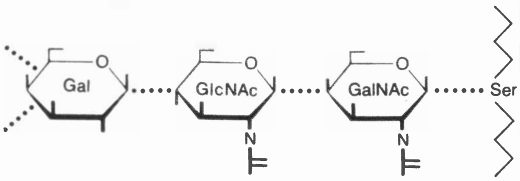
Because these enzymes do not trace a linear template the branching of oligosaccharides is readily accomplished. An enzyme can bind a sugar to the glycoprotein molecule wherever the specific reaction catalyzed by the enzyme is pos-

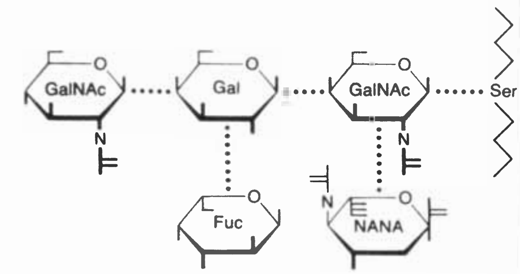


OLIGOSACCHARIDE SIDE CHAINS of glycoproteins form structures of great complexity. Because the carbohydrate chains are often branched, the number of possible structures for a given number of monosaccharides is much larger than it is for the same number of amino acids forming a polypeptide. Sugar units shown in gray are not always present. The blood-group glycoprotein shown is that of Type A blood; in Type B the acetylgalactosamine units shown in color are replaced by galactose. The broken lines con-

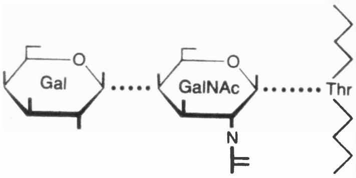
sible; it is constrained only by the availability of the substrates. Ultimately, of course, the saccharide side chains of glycoproteins are also expressions of genetic coding, but their structure is determined by the genes only at second hand. The primary products of the genes are the enzymes (which are themselves proteins); by controlling the synthesis of the enzymes the organism maintains indirect genetic control of saccharide structure as well.

The formation of the small carbohydrate side chains of collagen provides a relatively simple illustration of the proc-





PIG SUBMAXILLARY A MUCIN



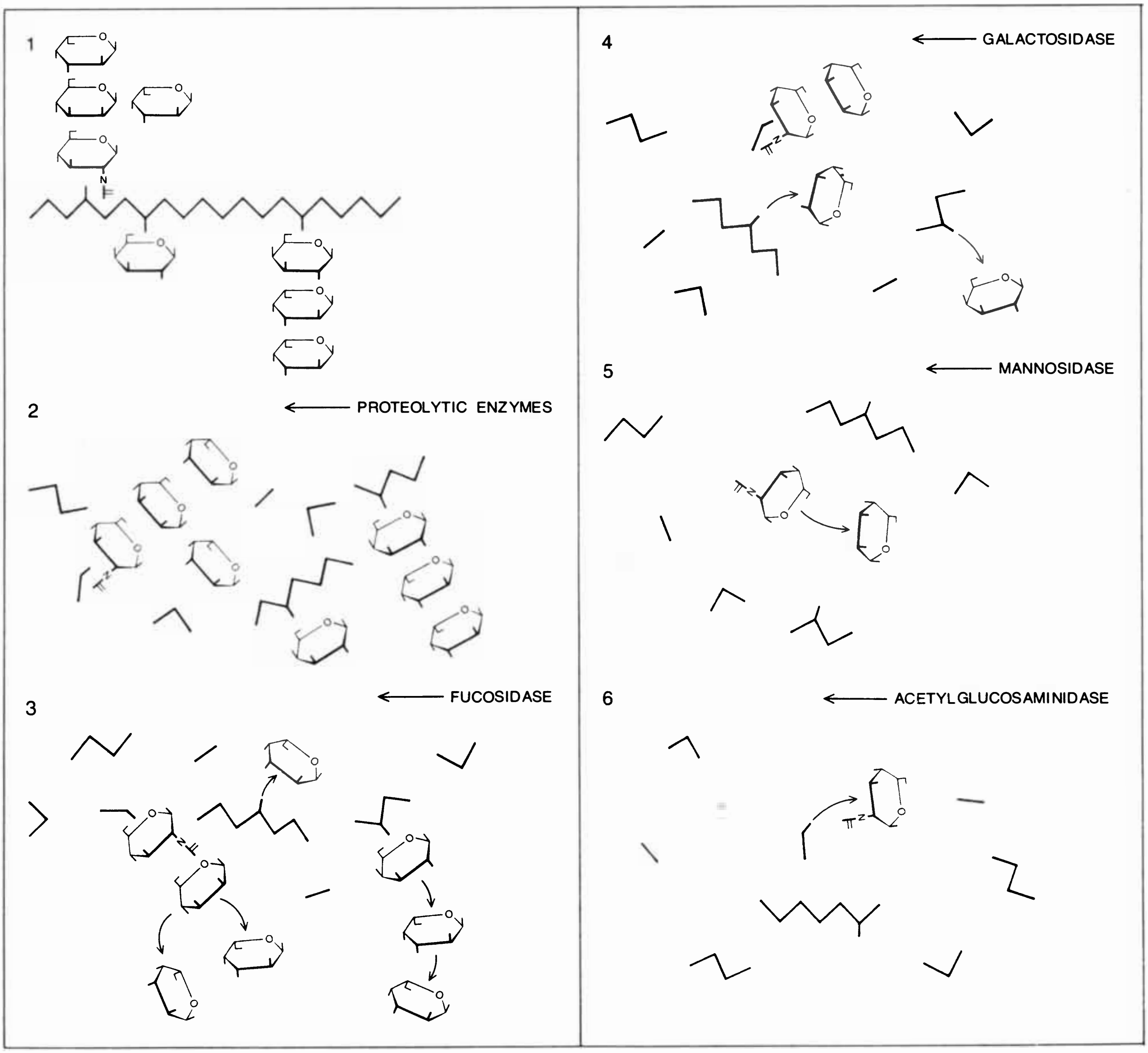
ANTIFREEZE GLYCOPROTEIN

necting monosaccharides in these drawings do not represent specific bonds; some of the structures, particularly the more complex ones, have been only tentatively determined.

ess of oligosaccharide synthesis. In this glycoprotein either a single galactose unit or the disaccharide glucosyl-galactose is linked to the amino acid unit hydroxylysine. The galactose is initially bound to a molecule of a nucleoside, uridine diphosphate (UDP); it is transferred to the hydroxyl group of hydroxylysine in a reaction catalyzed by one of a group of enzymes called glycosyltransferases. Each of these enzymes is specific to both the donor (in this case UDP-galactose) and the acceptor (hydroxylysine). In a second step a glucose unit is transferred from UDP-glucose to the bound galactose molecule by another glycosyltransferase of different specificity and cellular location. This enzyme will not attach glucose units to hydroxylysine or to free galactose but only to protein-linked galactosylhydroxylysine.

In spite of this specificity the glycosyltransferases are not as rigidly prescriptive as the genetic mechanism of protein synthesis. The reactions the enzymes mediate are influenced not only by the enzymes but also by environmental factors, such as the relative abundance of donor and acceptor molecules. Consequently the number of hydroxylysine residues in collagen that are unsubstituted, that carry only galactose and that carry glucosyl-galactose may vary. The result is a compound of heterogeneous composition.

A critical step in the synthesis of the carbohydrate side chains of glycoproteins is the attachment of the first sugar unit to the completed polypeptide chain. Studies of the amino acid sequence of isolated glycopeptides have



STRUCTURAL ANALYSIS of a glycoprotein begins with the digestion of the polypeptide chain (*1*) with proteolytic enzymes, yielding peptides to which carbohydrate side chains remain attached (*2*). These fragments are then treated with a succession of glycosidases, enzymes that cleave the glycosidic bond of specific monosaccharides and thereby remove sugars one at a time. Incubation with a fucosidase (*3*) removes the fucose units, which in this hypothetical glycoprotein occupy peripheral positions. In succeeding steps galactose (*4*) and mannose (*5*) are removed; finally acetylgalactosamine is released (*6*). Together with other methods of analysis this serial degradation can reveal both the structure of the side chains and the nature of the carbohydrate-protein linkage.

led to the identification of a linking region common to many glycoproteins. It consists of a tripeptide, asparagine-X-serine or asparagine-X-threonine, where X represents any amino acid. The saccharide associated with this tripeptide is acetylglucosamine, which bonds to the asparagine unit.

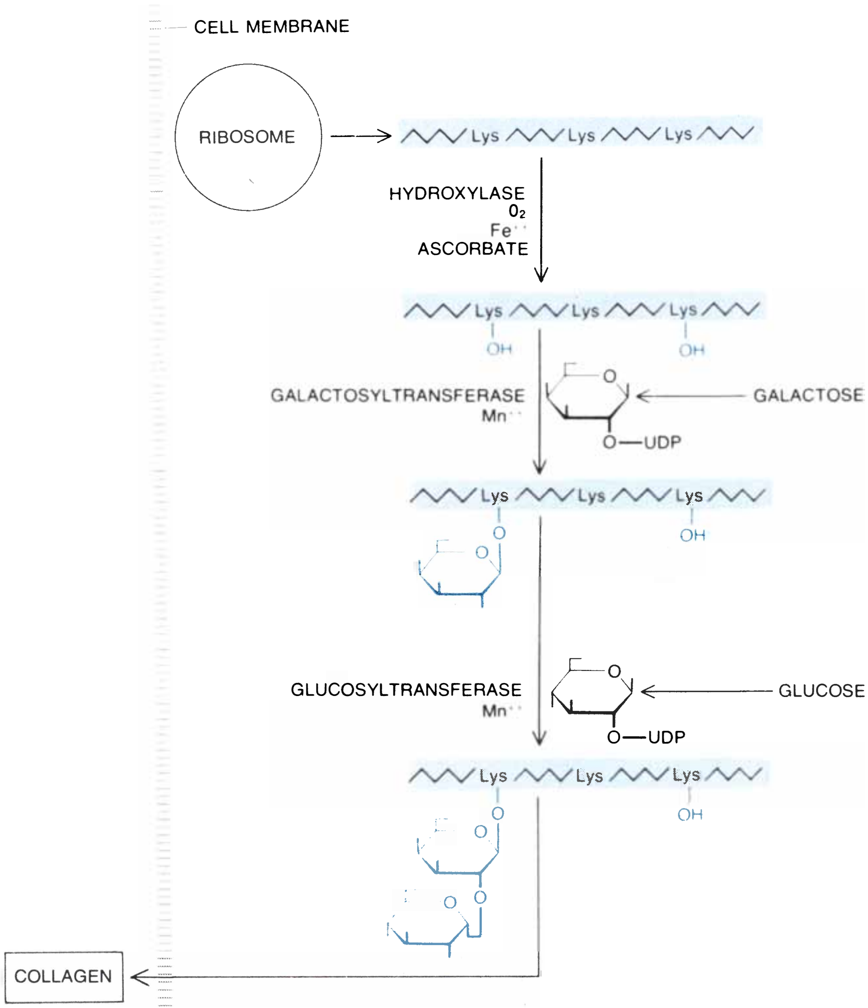
During the synthesis of certain glycoproteins this tripeptide may act as a marker recognized by the enzyme responsible for the attachment of acetylglucosamine. It is important to note, however, that not all such tripeptide sequences in proteins have sugars bound to them. Ribonuclease B, for example, has a carbohydrate side chain linked to the tripeptide asparagine-leucine-threonine; ribonuclease A, whose entire amino acid sequence is identical, has no carbohydrate component at all. It appears, therefore, that factors other than this sequence are involved in determining whether or not the initial sugar unit is attached.

The ultimate genetic control of glycoproteins is expressed with particular clarity in the human blood-group substances. The substances peculiar to both Type A and Type B blood are glycoproteins. The polypeptide component and most of each carbohydrate side chain of these molecules appear to be very similar; they differ only in a sugar that occupies a peripheral position. In Type A blood this sugar is acetylgalactosamine, in Type B it is galactose. The presence or absence of both peripheral sugars is determined by genetically specified enzymes (glycosyltransferases) that attach the sugars to an incomplete "core" of glycoprotein. People with Type A blood possess an enzyme that transfers acetylgalactosamine (from UDP-acetylgalactosamine) to the "core"; those with Type B blood have an enzyme that transfers galactose (from UDP-galactose) to the same core. People who lack both enzymes are Type O; those who have both belong to Type AB. The genes specifying the enzymes are inherited in a simple Mendelian manner.

Molecules that have been identified as glycoproteins perform multifarious roles in metabolism, and a few, such as the blood-group substances and the lectins, are of unknown function. Even where the function of the complete molecule is well understood, however, the role played by the carbohydrate portion often remains a mystery. Ribonuclease B, a glycoprotein, exhibits exactly the same enzymatic activity as ribonuclease A, which has no sugar. It therefore appears that the carbohydrate side chain does not affect the molecule's potency as an enzyme; if it does not, however, why is there sugar in ribonuclease B?

In a few cases the function of glycoprotein saccharides can be readily inferred. In salivary mucins, for example, the numerous sialic acid residues present on the exterior of the molecule repel one another and force the polypeptide chain to assume an extended rather than a globular conformation. Such extended molecules increase the viscosity of solutions, an important property in body fluids that serve as lubricants and protective agents. Special physical characteristics are also supplied by the carbohydrate component of the antifreeze glycoproteins of Antarctic fish; selective destruction or removal of the sugars abolishes the ability of the compounds to depress the freezing point of water solutions.

If a few sialic acid units are enzymatically removed from certain serum glycoproteins such as α_1 acid glycoprotein, the substances are rapidly eliminated from the circulation when they are injected into animals. The removal of sialic acid in these substances unmasks galactose units; if these too are removed, the rate of elimination is greatly decreased. G. Gilbert Ashwell of the National Institutes of Health and Anatol

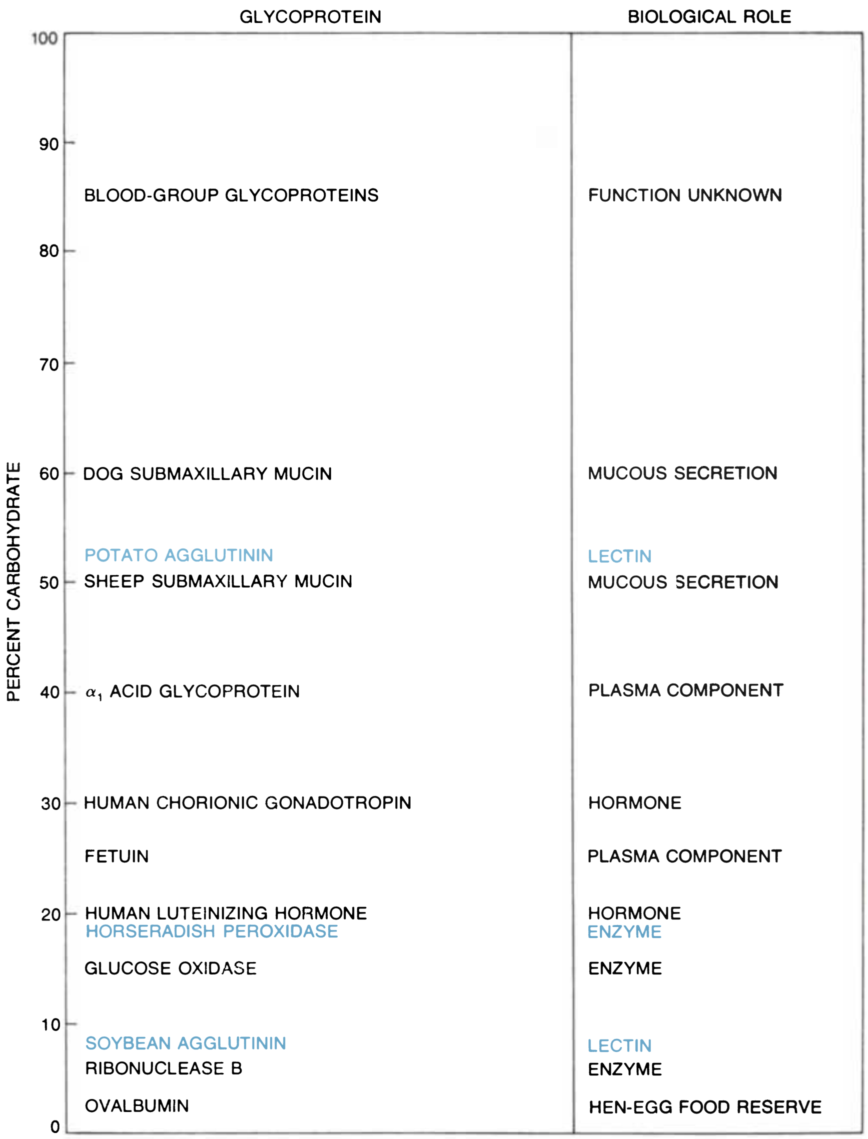


SYNTHESIS OF GLYCOPROTEINS in the cell takes place in two stages. The polypeptide portion of the molecule is assembled first; it is produced at a ribosome according to instructions embodied in the genetic material of the cell (DNA and RNA). The saccharide side chains are added in subsequent enzyme-mediated processes. In the synthesis of collagen, lysine is first converted to hydroxylysine by the addition of a hydroxyl group. A galactose unit is then transferred to this amino acid from a donor molecule, uridine diphosphate (UDP). This reaction is catalyzed by galactosyltransferase; the final step, in which a glucose unit is attached to the galactose, is controlled by glucosyltransferase. Various cofactors, such as iron and manganese, must also be present. Because the enzymatic process is not as rigidly prescribed as the genetic mechanism of polypeptide synthesis, some hydroxylysine residues may not acquire any carbohydrate and others may bear only galactose units.

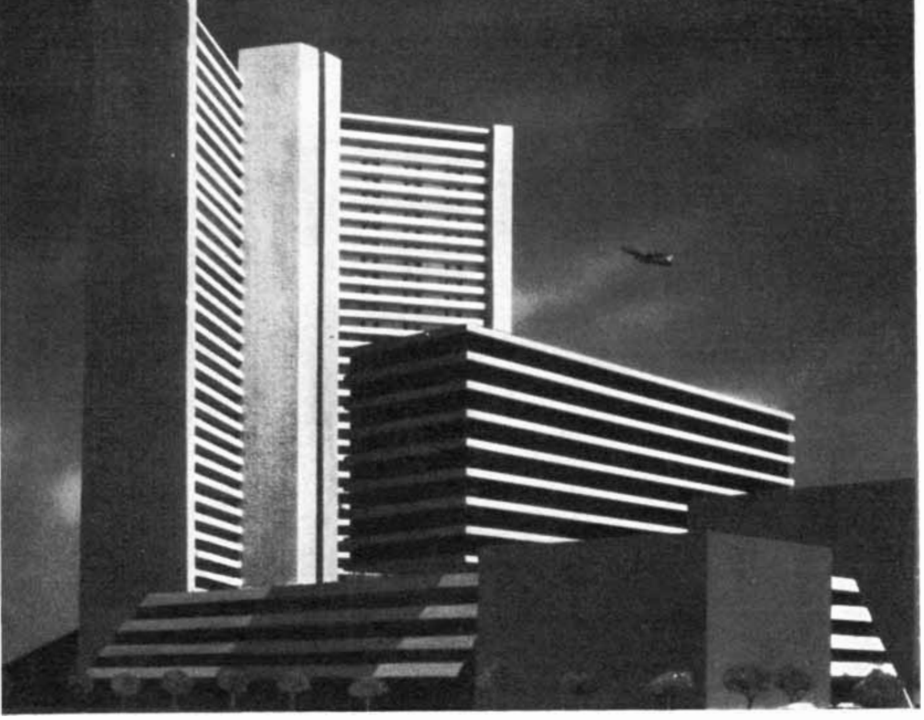
Morell of the Albert Einstein College of Medicine of Yeshiva University, who jointly discovered this phenomenon, have proposed that in some cases carbohydrate side chains may serve as markers or signals for the metabolic control of the protein by the organism. Other possible roles have been suggested: oligosaccharides may provide "handles," for example, for the transport of protein from one part of the cell to another, or they may render the protein resistant to enzymatic digestion.

The most intriguing of these proposals, however, is that the sugars of glycoproteins may be the essential component of the mechanism by which cells recognize other cells and large molecules. These sugars, perched on the surface of the cell, appear to act as letters or words in intercellular communication.

The importance of cellular communication is perhaps best illustrated by its role in the development of cancerous growths. The social behavior of malignant cells is different from that of normal cells. Normal cells stop growing when they touch each other, for example, whereas cancer cells grow without restraint. Much evidence suggests that the transformation that makes malignant cells unresponsive to such environmental cues is related to a modification of the cell surface. Recent investigations, some of which have been conducted in our laboratory, strongly suggest that this modification of the surface properties of the cell involves the sugar side chains of glycoproteins.



CARBOHYDRATE CONTENT of plant (*color*) and animal (*black*) glycoproteins ranges from about 1 percent to more than 85 percent. There are also variations in the size of the saccharide side chains, in the number of side chains per molecule and in molecular weight.

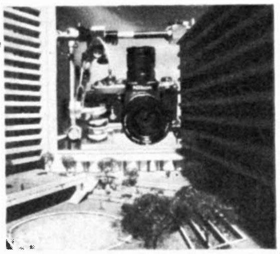


Architects Langdon & Wilson, A.I.A.

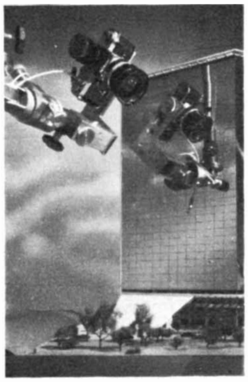
This building is 2 feet high

It's a model, of course. But, a model so well detailed and so superbly photographed that you can almost hear the hum of the elevators. If you were a prospective tenant, you could readily visualize its effect on the prestige of your business.

In achieving this exceptional [illegible], architectural photographer Jack Laxer has demonstrated for us not only his expertise but also his innovative and highly imaginative approach to this demanding application. When he recently acquired his new Nikon F2 cameras, he went to extraordinary lengths in exploiting their capabilities.



He designed ingenious support systems that allow totally flexible camera movement in all three directions and provides precise adjustments for about 800 different camera positions. It was a sizable investment in time, effort and money—but the unique combination of cameras, customized support system and talent has already paid off. The architects and builders, who are Laxer's clients, now can count on greater realism in photography for their promotional brochures. In fact, his Nikon photos are so accurate and so detailed that they enable clients to make design changes before the actual construction begins.



Apart from its legendary precision and reliability, the Nikon F2 offers special advantages for this type of work. Some are built in, such as the viewing system, the most accurate among 35mm cameras, and the multi-exposure provision that lets you add a background or put clouds in the sky.

Others are available in the Nikon System: The perspective-correcting PC-Nikkor lens, actually designed for architectural work. The 6X Magnifying Finder providing, as Laxer puts it, "an image virtually equal to 6x9 inches" The unique Nikon Bellow Attachment IV whose front standard movements permit corrective closeup photography. And, two of the nineteen interchangeable finder screens have grid patterns that simplify alignment of vertical and horizontal shapes in critical compositions.

This kind of precision photography has always been considered the exclusive domain of large, heavy studio view cameras. Yet, here it is done professionally with a 35mm camera that is best known for its speed and agility in handling split-second news, sports and similar action.

But, that is what Nikon versatility is all about. To help you make superior photographs of whatever interests you, whether a model just inches from the camera or a solar eclipse across millions of miles of space. To inspire you to look for new visions, new ways to picture your world and to test the capabilities of a camera whose limits have yet to be reached. And, to help you put it all together, there's the Nikon School of Photography, traveling to all parts of the U.S.A.

Your Nikon dealer will be glad to demonstrate the Nikon F2 and its system. Ask also about the Nikon School. Or write for folio 19, Nikon Inc., Garden City, New York 11530. Subsidiary/Ehrenreich Photo-Optical Industries, Inc. EPOI (In Canada: Anglo-Photo, Ltd., P.Q.)

THE PARTICLES OF WEAR

When the parts of a machine wear, they shed tiny bits of metal. A new method of analyzing such particles clarifies the process of wear and can give early warning of the failure of a machine

by Douglas Scott, William W. Seifert and Vernon C. Westcott

In a machine part such as a bearing wear is the removal and deformation of surface material. The process, which deposits tiny particles of metal in the lubricant of the machine, can be slow and normal or fast and catastrophic. Distinguishing between the two states is obviously a matter of concern to the mechanical engineer, and a strong clue is provided by the particles of wear. A new method of detecting and isolating such particles both reveals abnormal wear and elucidates the entire process.

Some machines built in the 19th century operated for as long as 70 years without failure. Much of their longevity was the result of continuous maintenance. On the walking-beam steamboat, for example, an oiler frequently inspected and lubricated the engine. This specialist could see most of the moving parts, and if one began to wear rapidly, remedial action could be taken before the machine failed.

Machines today are compact and the moving parts are totally enclosed. In order to determine the type and rate of wear the modern maintenance engineer must periodically dismantle the equipment. Such a procedure not only is costly but also may decrease rather than increase the length of time that the machine will operate without failure. The danger of failure is much greater at startup than after the machine has been running for some time.

In 1970 two of us (Seifert and Westcott) began to look for a simple method of detecting the incipient failure of machine parts. At that time the emission spectrograph had come into fairly wide use for detecting abnormal wear in machines. The method is based on the fact that a rapid increase in the amount of metal in the oil indicates that a part is wearing abnormally. Although the emission spectrograph can measure the quantity of each of several metals in lubricating oil, it does not distinguish between particles of free metal and particles of metal oxides (or other compounds of the metal).

After examining the problem we concluded that it would be necessary to find out what types of metal particles were in the oil before further progress could be made. We collected samples of oil from airplane engines and examined them in a high-quality transmitted-light microscope. Although we could see particles in the oil, we quickly realized that the number of particles did not correspond to the amount of wear metal in the oil as determined with the spectrograph. Next we examined samples of oil in an ultramicroscope. In this instrument the specimen is illuminated at an angle, so that if a clear liquid is examined, no light is seen, but if particles are present, the light is scattered and enters the objective lens. Particles much smaller than a wavelength of light can be detected.

The oil samples were found to contain a large number of wear particles ranging in size from several micrometers (thousandths of a millimeter) to as small as 20 nanometers (millionths of a millimeter). There were about a trillion (10^{12}) particles per cubic centimeter of oil. In addition to the wear particles there was a larger number of other kinds of particles that were not the result of wear. They came from the air and other sources of contamination.

Since most of the wearing parts in a machine are made of steel, we thought that a magnet would attract the steel wear particles and that the resulting motion would be visible in the ultramicroscope. To our surprise, when a powerful Alnico magnet was put on the microscope stage, none of the particles in the field moved. This prompted a theoretical analysis of the magnetic forces that would be exerted on tiny steel particles by such a magnet. The conclusions drawn from the analysis confirmed the observed fact that the attractive force is small.

The typical wear particles are so minute that their north and south magnetic poles are very close together. When they are put in a magnetic field, the attraction of one pole is almost perfectly balanced by the repulsion of the other. The particle immediately rotates to align its poles with the magnetic field, but it is not induced by the small residual force to move through the viscous oil.

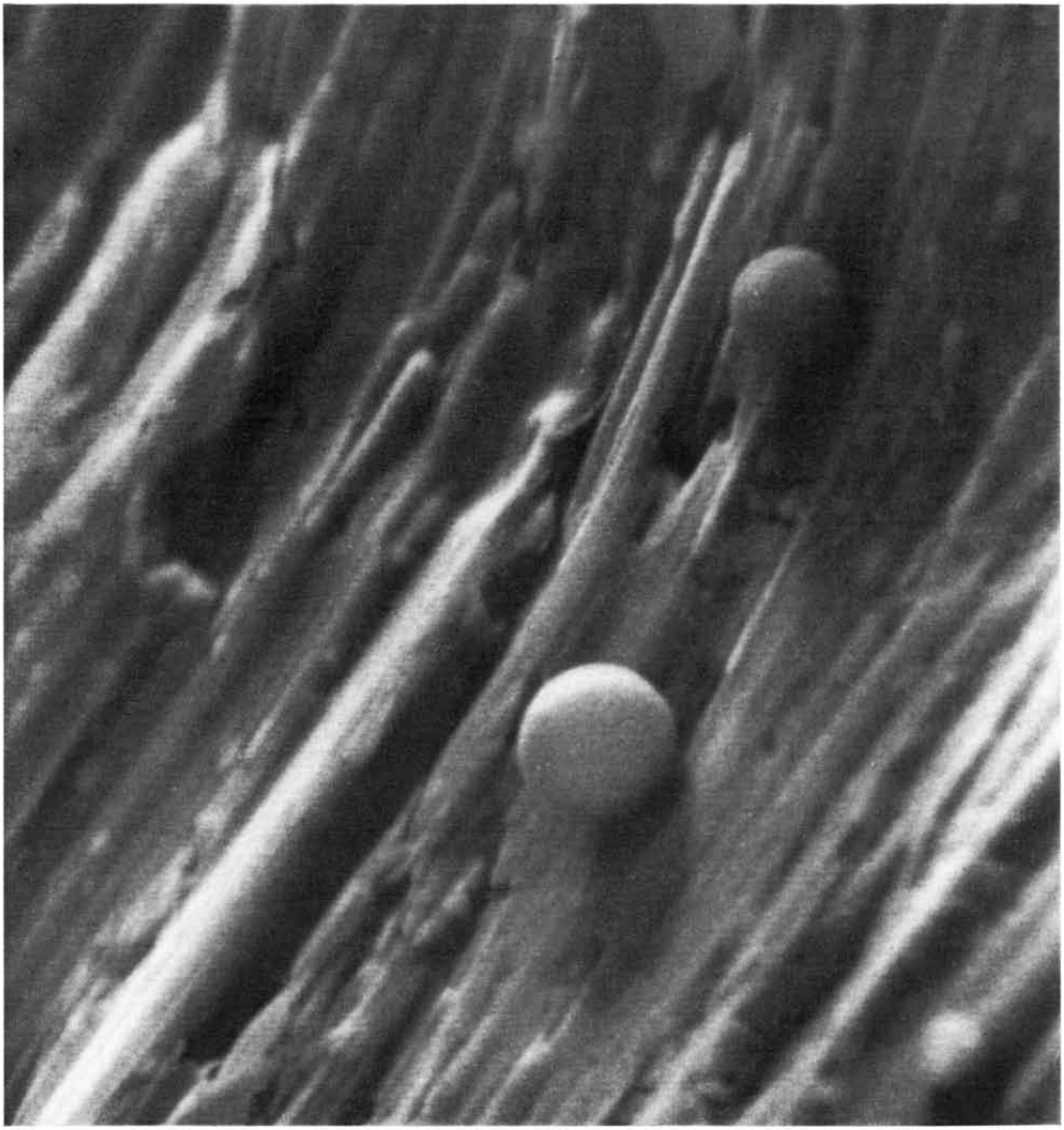
Next we placed a specially designed magnet with a large field gradient on the microscope stage. This magnet attracted a small fraction of the particles in the oil. The reason is that in a highly divergent magnetic field the force acting on one pole of a particle is substantially different from the force acting on the other pole. Further studies showed that most of the very small metal wear particles behave like permanent magnets. The particles are so small that they have only one magnetic "domain," or at most a few domains. Indeed, they are among the most powerful magnets in existence. The larger particles of steel in the oil have thousands of randomly oriented magnetic domains and do not exhibit permanent magnetism.

The examination of oil from jet engines that had failed showed that although it contained a higher concentration of steel particles, they had a lower magnetic moment than the particles in oil from engines that were running satisfactorily. This finding indicated that when an engine began to fail, the average size of the wear particles in the oil increased. We concluded that it would be worthwhile to isolate the wear particles from the oil and determine their size distribution.

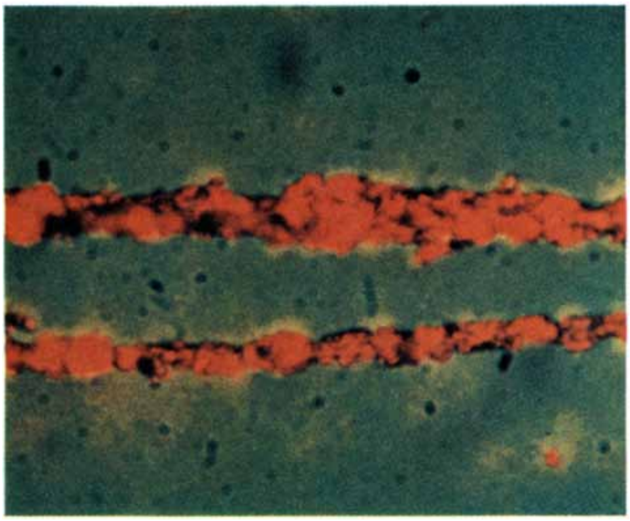
Filters with specific pore sizes make it possible to recover most of the particles in the oil, but the filters do not distinguish between the wear particles and the nonwear ones. The piles of assorted particles that are collected by the filters are thus essentially "dirt." As a result of the discovery of the unusual magnetic properties of wear particles, however, we developed a new instrument that has come to be known as the Ferrograph analyzer. In this instrument the sample of oil to be examined passes through a highly divergent magnetic field as it is running down a microscope slide with a specially treated surface.

After the oil has been removed from the slide the particles are found arranged according to size along the length of the slide. Since the wear particles are precipitated magnetically, virtually all the unwanted dirt particles are eliminated, and since the precipitated particles are deposited according to size, they can be viewed easily. Large particles are not obscured by clusters of small ones, and the unique characteristics of the small particles can be determined.

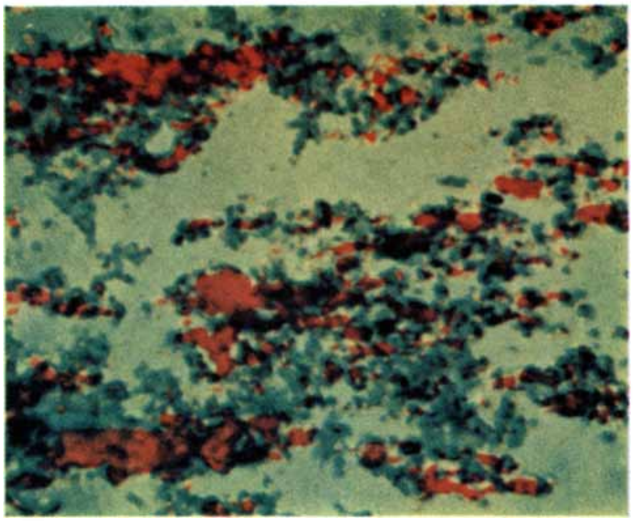
The Ferrograph analyzer consists of a pump that delivers the oil sample at a low rate (on the order of .25 cubic centimeter per minute), a magnet that de-



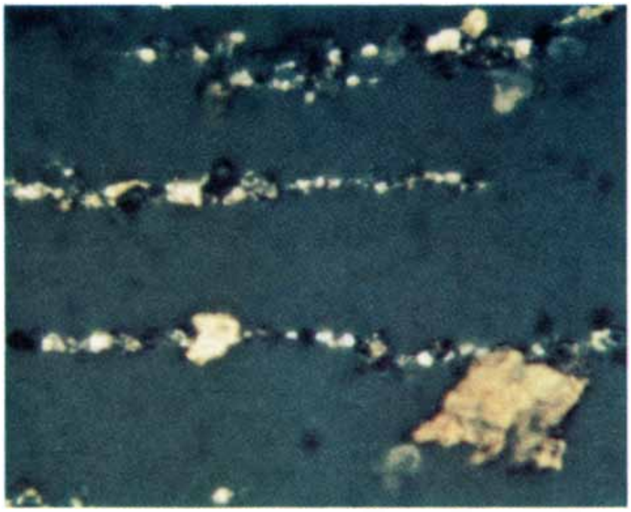
SPHERICAL WEAR PARTICLES were discovered in microscopic fatigue cracks in the balls of ball bearings when they were examined in the scanning electron microscope. Two spheres on the surface of a microcrack are shown here enlarged 5,000 diameters.



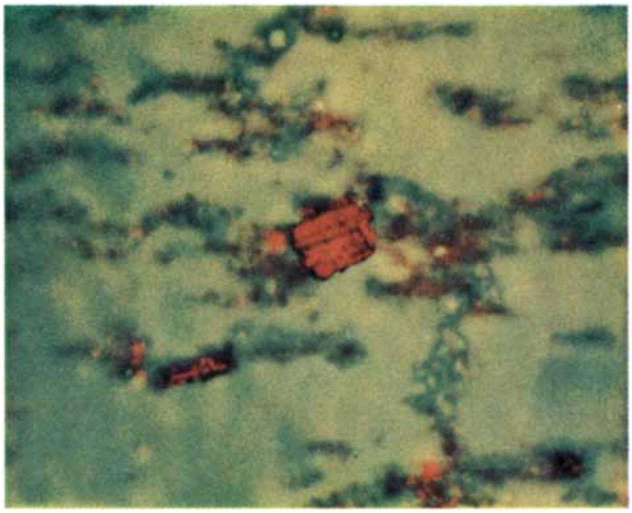
RUBBING-WEAR PARTICLES from cold-rolled steel, seen in the bichromatic microscope (*see illustration on opposite page*), are enlarged 1,300 diameters. Strings follow magnetic lines of force.



STAINLESS-STEEL WEAR PARTICLES, enlarged 1,300 diameters, form short, fat strings that are readily distinguished from the long, thin strings formed by the wear particles of low-alloy steels.



BRONZE RUBBING-WEAR PARTICLES, enlarged 900 diameters, are weakly magnetic and appear along with smaller steel particles.



FATIGUE PARTICLE, enlarged 900 diameters, from lubricating oil of helicopter transmission exhibits fractured crystal faces.

CUTTING-WEAR PARTICLE, enlarged 1,300 diameters, shows colors caused by heating. Note small sphere to the right of center.

SPIRAL CUTTING-WEAR PARTICLE, very similar to chips produced by a metal-cutting lathe, is shown enlarged 1,600 diameters.

velops a high-gradient magnetic field near its poles and a treated microscope slide. The magnetic particles adhere to a substrate on the slide, which is mounted at a slight angle to the horizontal. The oil sample is first diluted with a special solvent in order to increase the mobility of the wear particles in it. The diluted oil is pumped to the higher end of the slide and flows off the lower end [*see top illustration on next page*].

The magnetic field is stronger at the exit end than at the delivery end. As a result the particles are subjected to a continuously increasing magnetic force as they flow along the slide. In the presence of the strong magnetic field of the Ferrograph analyzer all the steel particles, even the largest, are magnetically saturated, so that the attractive force is proportional to the volume of the particle. The combination of drag forces and magnetic forces causes large particles to be deposited first. As a result of the higher surface-to-volume ratio of the smaller particles, they migrate through the oil slowly and flow with the oil for a greater distance before they are deposited.

After approximately two cubic centimeters of the sample has been pumped across the slide, a washing cycle removes the oil and a fixing cycle causes the wear particles to adhere to the substrate. The liquids used for these purposes evaporate quickly and leave the particles permanently attached. A normal pumping time for an oil sample is five minutes, with another three minutes required for washing and fixing.

The deposited particles appear as a dark band along the slide [*see middle illustration on next page*]. By measuring the optical density at various points on the band one can determine the average density and size distribution of the wear particles. We have found that a rapid increase in the ratio of large particles to small ones indicates that severe wear has started. Although valuable information is obtained from the optical-density readings, much additional information about the nature of wear particles has been gained from microscopic examination of the particles on the slide.

When a slide is viewed in a microscope using reflected light, it is difficult to distinguish the free metal from the compounds. In the transmitted-light microscope the particles appear as black shadows with no surface detail. In order to examine the particles more effectively we had to develop a new technique of microscope illumination, one in which both reflected and transmitted light are used but in which the reflected light is red and the transmitted light is green. The instrument is called the bichromatic microscope [*see illustration at right*].

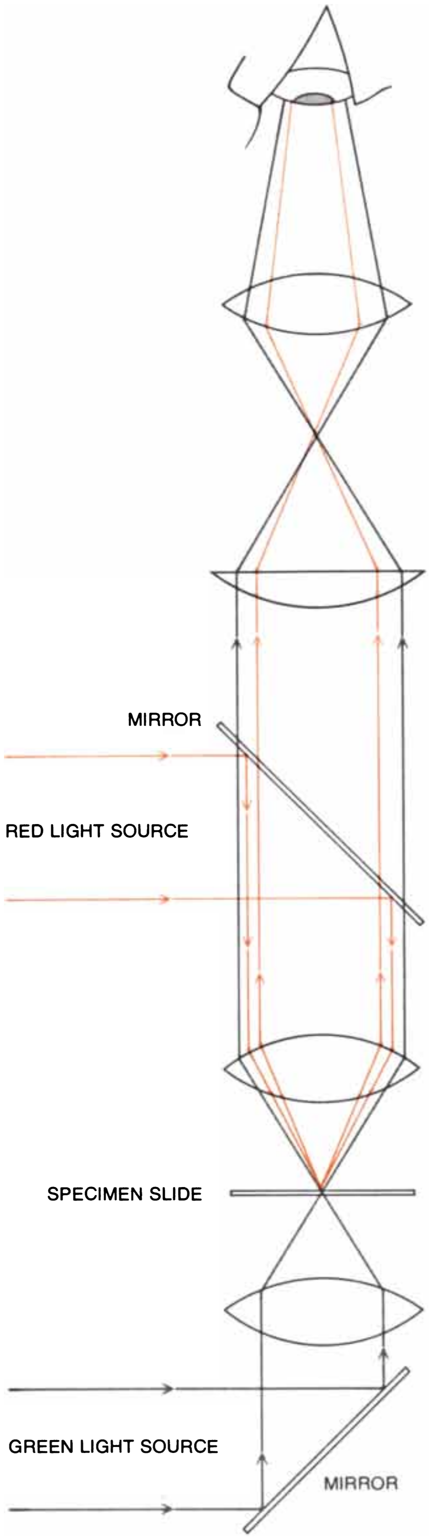
This lighting arrangement exploits the fact that metals attenuate light much more than compounds do. That is, metals contain free electrons that absorb and reflect light waves after they have penetrated only a few layers of atoms. Particles consisting of compounds allow much of the light to pass through them. The result in the bichromatic microscope is that metal particles as small as one micrometer will reflect the red light while blocking the green light and therefore appear to be red. Compound particles that allow much of the green light to pass through them will appear to be green. If they are several micrometers thick, they will be yellow or pink. Metal particles do not have to be highly reflecting in order to be visible, and particles with different surface conditions are distinguishable by the intensity of the red light they reflect.

Extensive microscopic examination of particles deposited on Ferrogram slides has revealed that wear particles have characteristic physical properties such as shape and composition. Moreover, particles generated by different kinds of wearing mechanisms can be distinguished from one another. In these studies we were greatly assisted by colleagues at the National Bureau of Standards and in the Department of Mechanical Engineering at the Massachusetts Institute of Technology; we also received much help from Richard S. Miller of the Office of Naval Research.

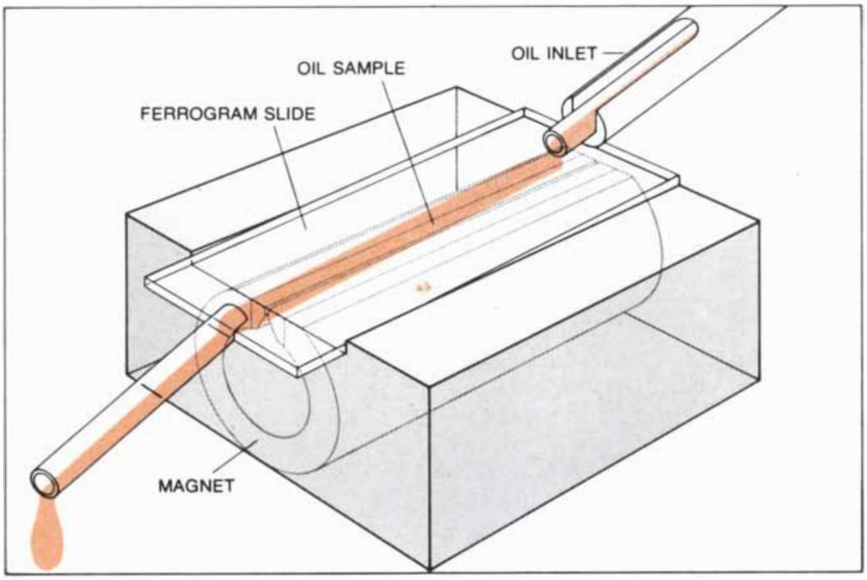
The wear particle most frequently observed is generated by rubbing or adhesive wear. These particles are asymmetrical and can be as long as five micrometers or as short as 20 nanometers. They are found in the lubricants of all machines and can be considered to be indicators of normal wear. They are produced in large quantities when one metal surface slides over another, for example when a piston moves in a cylinder of an automobile engine.

The classical explanation of the process by which rubbing-wear particles are formed assumes that contact between metal surfaces occurs where asperities, or microscopic high spots, on one surface touch and adhere to asperities on the opposing surface. As the surfaces move over each other some of the asperities presumably shear off and become welded to the opposing surface. In some instances asperities break off completely and enter the lubricating oil.

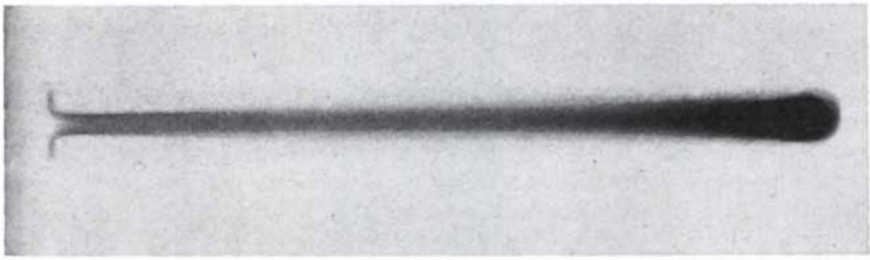
Our early work with wear particles isolated on Ferrogram slides revealed



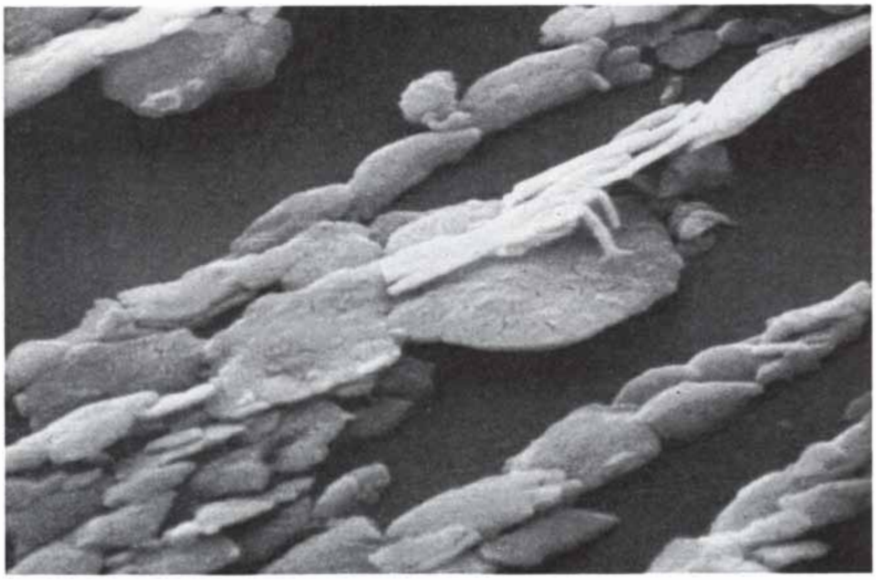
BICHROMATIC MICROSCOPE has two light sources, one green and the other red. The green light is transmitted through the specimen and the red light is reflected from its surface. This arrangement makes it possible to distinguish between free metal particles and oxides or other compounds. Free metal particles as small as one micrometer (thousandths of a millimeter) reflect red light and are seen as red. Compound particles let green light through and depending on thickness appear green, yellow or pink.



"FERROGRAPH" ANALYZER separates wear particles from lubricating oil by precipitation in a strong magnetic field. Larger particles are deposited near the oil inlet; smaller magnetic particles migrate farther along the slide. After about two cubic centimeters of the oil sample has passed over the slide, washing and fixing solutions are pumped over the substrate of the slide to remove the remaining oil and to bond the particles to the surface.



DEPOSITED WEAR PARTICLES on a Ferrogram slide form a dark band with large particles at the inlet (*left*) and small particles at the outlet (*right*). The size distribution of the particles can be determined by measuring the optical density at various points on the band. Particles can be examined in a microscope to determine the type of wear taking place.



STRINGS OF RUBBING-WEAR PARTICLES on a Ferrogram slide are enlarged 6,500 diameters in this scanning electron micrograph. The disk-shaped particle with faint radial marks probably was formed by being squeezed between two hard opposing surfaces.

that the classical explanation was in error. The rubbing-wear particles appear as thin flakes of metal with highly polished surfaces [*see bottom illustration at left*]. Moreover, when the polished surface of a machine part is examined with the scanning electron microscope, it appears to be extraordinarily smooth.

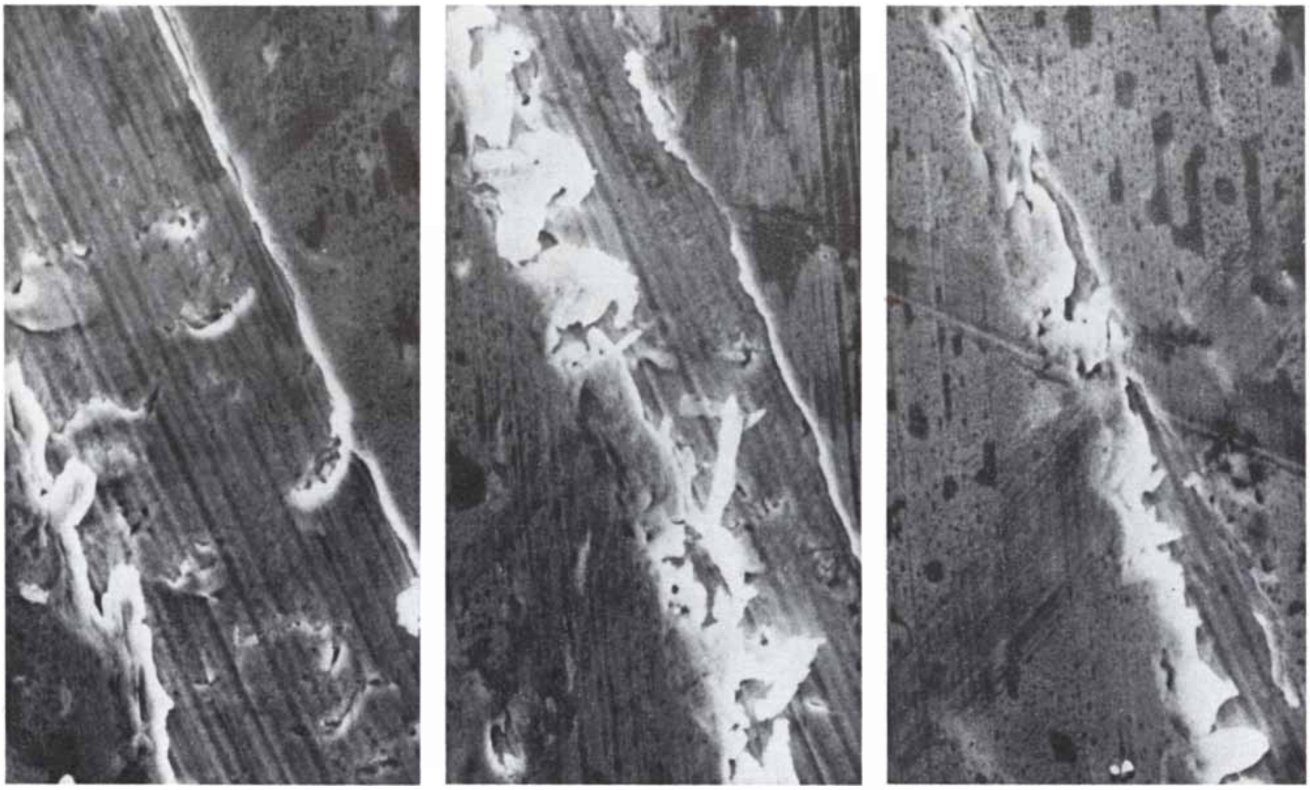
The rubbing of opposing surfaces polishes the surface and creates an upper layer that is different from the underlying metal. This layer is very similar to a layer on polished metal surfaces that was first noted by Sir George Beilby some 70 years ago and is now named after him. The Beilby layer does not require direct metal-to-metal contact for its generation, although it is usually formed that way. Beilby was able to produce the layer by polishing metal with a chamois skin. Stropping a razor is another way of producing a Beilby layer.

The Beilby layer has the same constituents as the parent metal but the long-range order of the metal crystals is missing. It has largely been neglected by modern metallurgists because they routinely etch metal specimens before examining them, and etching dissolves the Beilby layer. The material of the layer is almost superductile and spreads over cracks and other irregularities in the underlying metal. Repeated rubbing causes the Beilby layer to become fatigued, and particles of it flake off [*see bottom illustration on opposite page*].

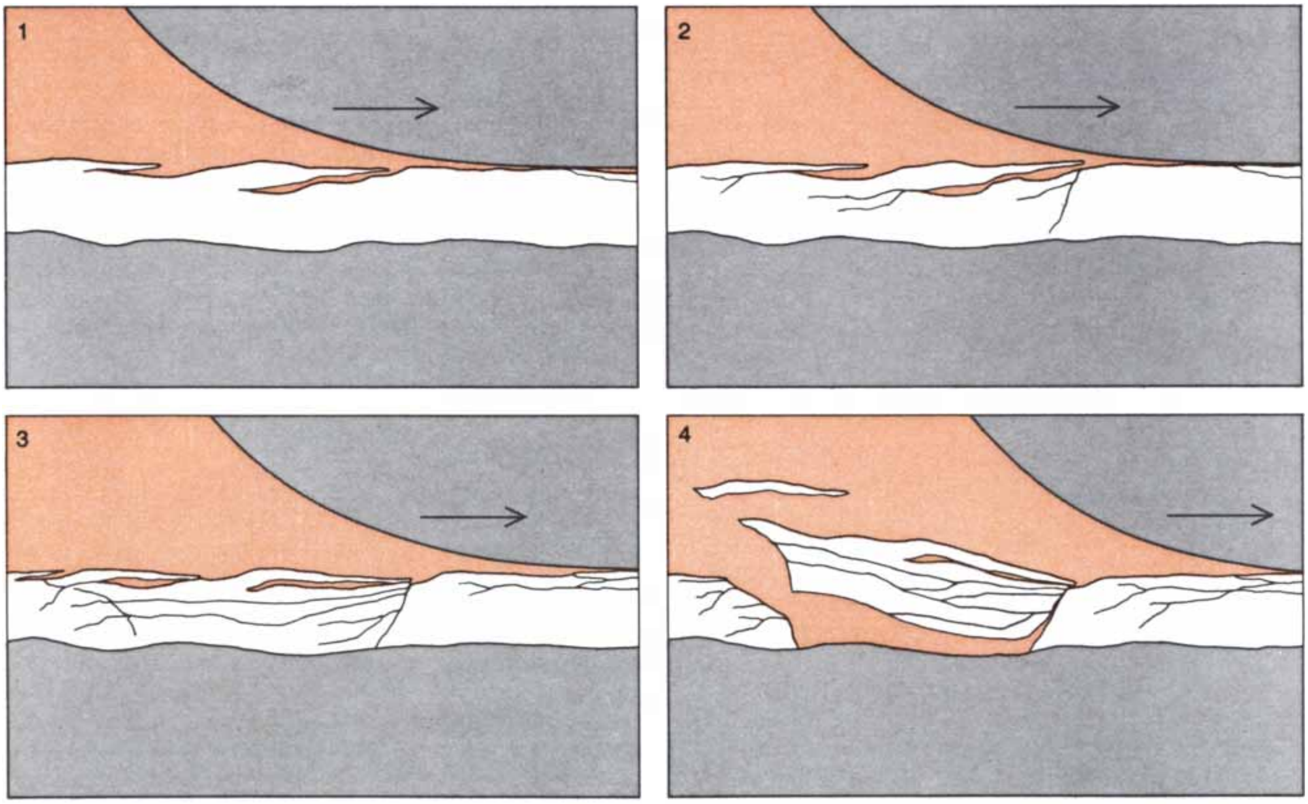
On the Ferrogram slide rubbing-wear particles are marshaled into chains by the lines of force in the magnetic field. Many of these chains include other particles. They may be flakes of iron oxide from the oxidized surface of the steel or oxide inclusions present before oxidation. Among the wear particles of hard steels are refractory, pebble-shaped particles consisting of an iron oxide (Fe_3O_4) with iron dissolved in it. Their presence points to the class of steel being worn.

Some machine parts are made of nonmagnetic stainless steel. Even these parts, however, can produce wear particles that are magnetic. The effect is attributed to cold-working, which converts austenite, a nonmagnetic solid solution of iron and carbon, into martensite, a magnetic solid solution. Particles from stainless-steel parts tend to deposit in short, fat strings that are easy to distinguish from the long, thin strings formed by particles from low-alloy steels.

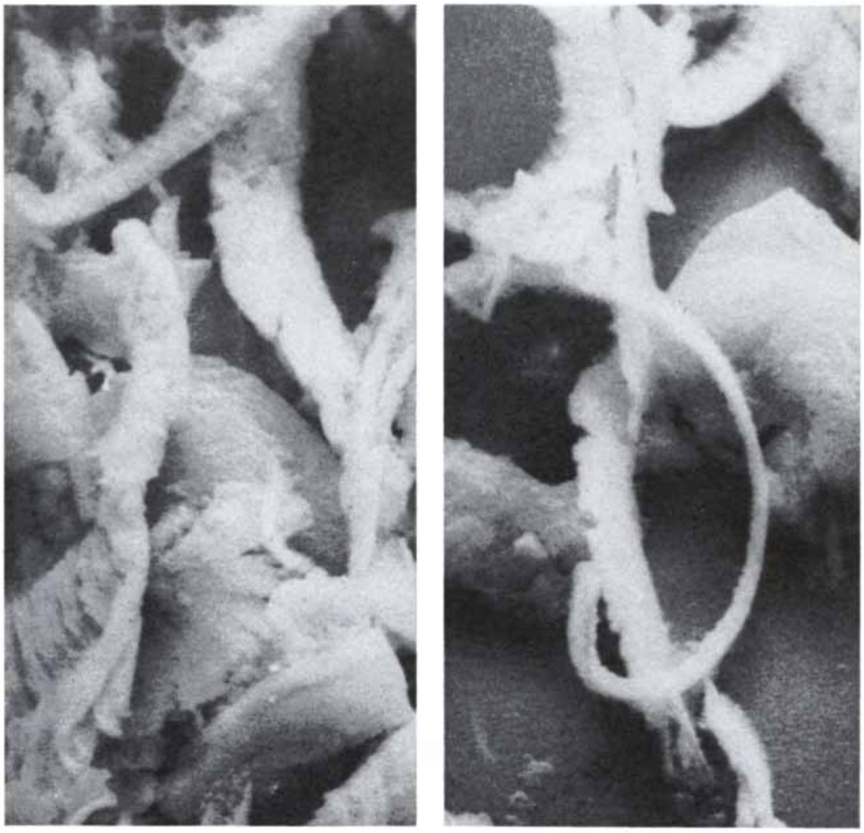
Rubbing-wear particles of an unusual class are generated when a nonferrous metal is in contact with a ferrous one. The nonferrous particles that are formed are slightly magnetic and are deposited



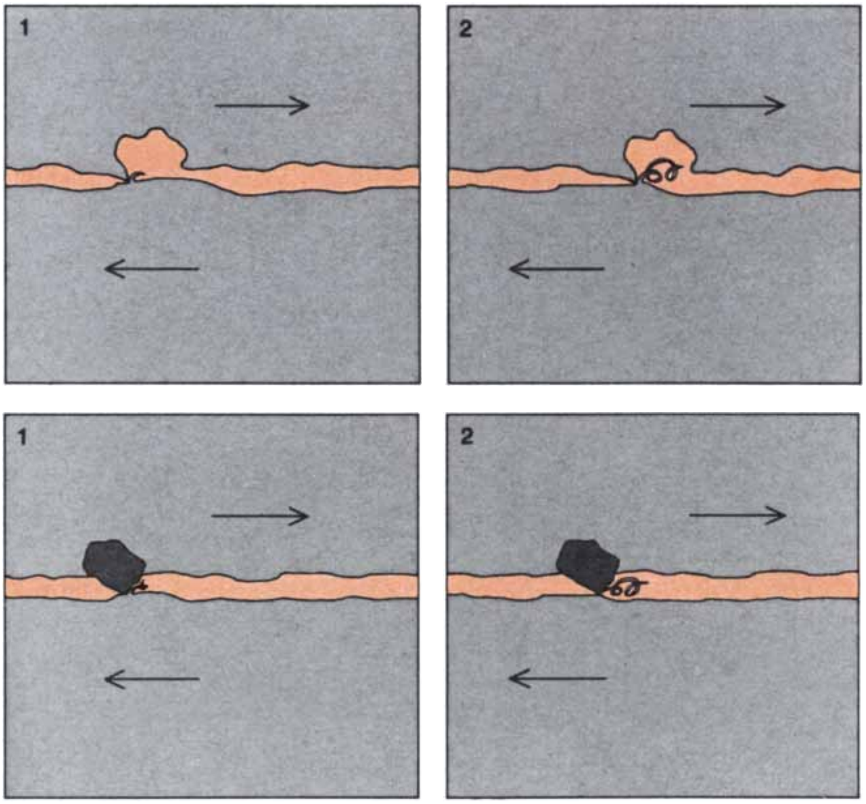
BEILBY LAYER is formed on a machined bronze surface, enlarged 1,100 diameters, as a piece of steel is rubbed over it (*left*). The motion of the steel is from lower left to upper right and at right angles to the wide machining scratch. As the rubbing continues the Beilby layer covers the scratch (*center and right*). Rubbing-wear particles break off. These scanning electron micrographs, provided by Harry Koba of the Massachusetts Institute of Technology, do not show identical portions of the bronze surface.



RUBBING WEAR is the most common kind of wear and occurs when one metal surface slides over another metal surface. The repeated rubbing produces a thin Beilby layer, shown in white, in which the long-range crystalline order of the atoms of the metal is disrupted. Often the Beilby layer is not tightly bonded to the underlying surface, and with successive passes of the metal surfaces, sections of the layer spall off. In addition smaller portions of the layer may flake off. Only wear of the lower metal surface is shown.



CUTTING-WEAR PARTICLES formed after the initial failure of two main bearings in a jet engine are enlarged 2,000 diameters in these scanning electron micrographs. The particles were collected on a Ferrogram slide. Many show serrations or chatter marks.



CUTTING WEAR occurs when particles are machined from a metal surface either by a sharp projection (*top*) or by an abrasive particle (*bottom*). A rapid increase in the number of cutting-wear particles in the lubricating oil indicates machine failure may be imminent.

farther along the Ferrogram slide than might be expected on the basis of their size [*see middle illustrations on page 90*]. It is believed that in the wear process minute magnetic steel particles become embedded in the nonferrous particles.

A different type of particle is produced by cutting, or abrasive, wear. Such particles are manufactured when an abrasive particle on one surface cuts strips from the opposing surface. They look like chips from a metalworking lathe, and they take the form of spirals, loops and bent wires.

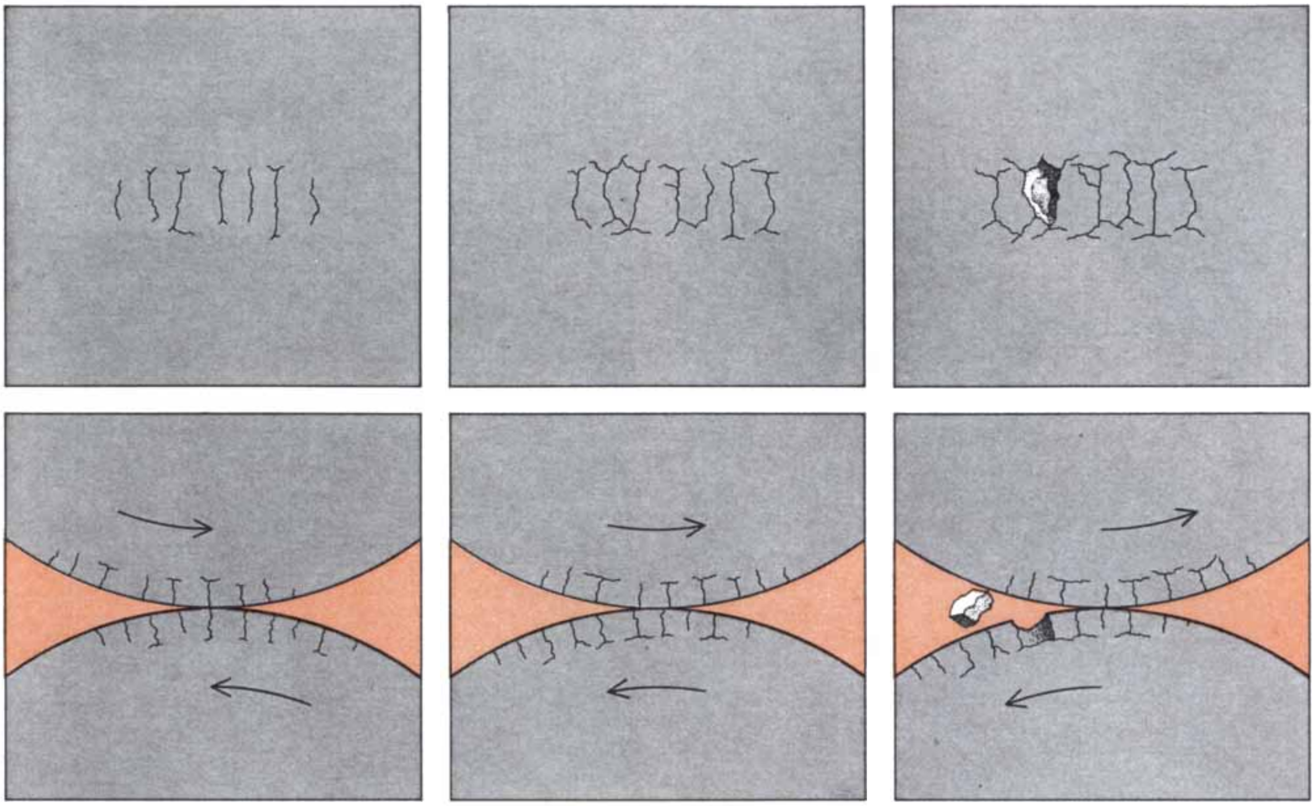
Cutting-wear particles can be generated without the presence of an abrasive. Under some conditions sharp projections on the surface of a machine part can give rise to cutting-wear particles. On the other hand, the presence of an abrasive such as sand in the lubricating oil does not necessarily lead to the generation of cutting-wear particles. It may be that the sand is not sufficiently hard or that the sand particles are not held in a position where they can cut chips from a metal surface.

The presence of individual cutting-wear particles is not significant, but the presence of several hundred is an indication that a severe cutting-wear process is under way. A sudden increase in cutting-wear particles in successive oil samples indicates that machine failure is imminent.

Surface fatigue resulting from very high stresses generally gives rise to characteristic chunklike particles in which the three perpendicular dimensions are roughly the same. For example, in a highly loaded gear system the surface of a gear may develop a series of minute cracks. In the scanning electron microscope such a surface looks much like the clay bottom of a dried-up reservoir. Particles are spalled from the surface by the repeated buildup and release of pressure.

In the course of examining scanning electron micrographs of ball bearings that had failed one of us (Scott) and his colleagues at the National Engineering Laboratory in Glasgow discovered minute spherical particles in the fatigue cracks. At about the same time tiny spherical particles were found on some Ferrogram slides. Comparison of photographs taken by the two groups showed that the spheres appeared to be identical. How these spheres come to be formed has turned out to be one of the most interesting aspects of our investigation of wear particles.

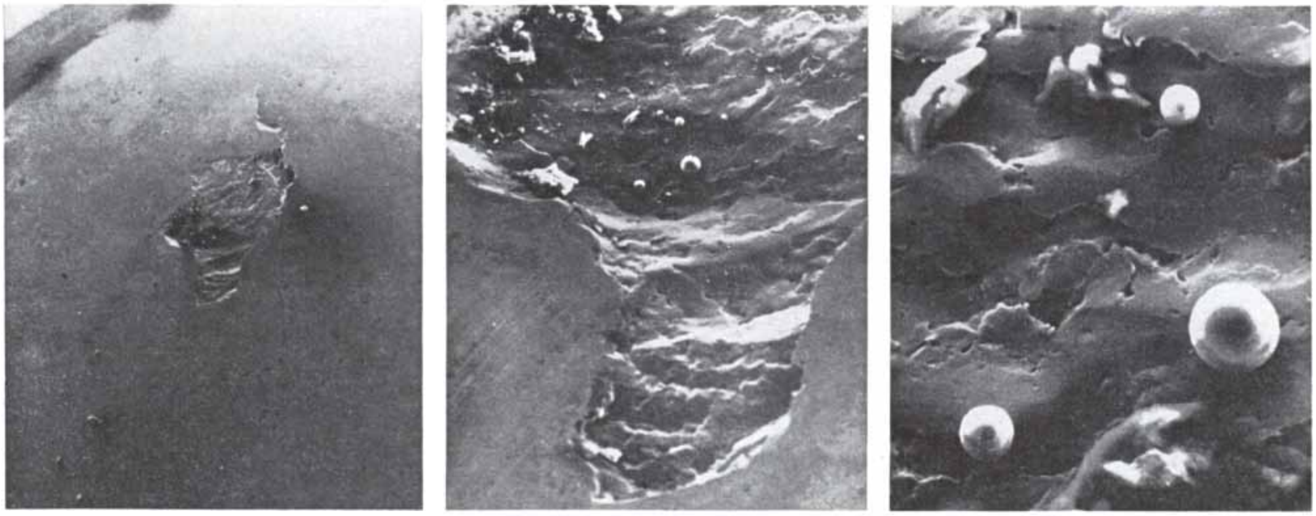
A fatigue crack in a ball bearing or in



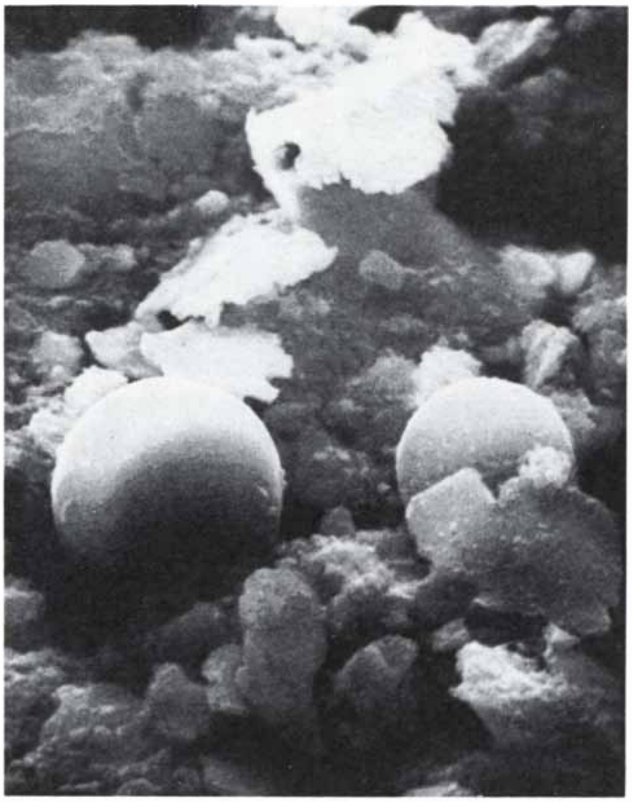
SURFACE-FATIGUE WEAR results from the high stresses imposed by repeatedly loading and unloading, for example near the tip of a heavily loaded gear tooth. Cracks develop on the surface (*top illustrations, showing top view*) and perpendicularly to it (*bottom illustrations, showing cross-sectional view*). Eventually the surface and underlying cracks link up and rectangular chunks spall out, leaving minute pits on the surface. The arrows indicate the direction of the reciprocating sliding motion of the gear teeth.

its race may either begin at the surface and propagate inward or begin below the surface and propagate outward. Modern steels contain few foreign inclusions, so that the probability of a subsurface flaw is low. It is therefore believed many of the cracks originate at the bearing surface. They propagate into the material at a small angle with respect to the surface and then turn parallel to the surface.

The pressure directly under the contact point of a bearing can reach 40,000 kilograms per square centimeter in a heavily loaded system. As the balls roll over the race the minute cracks that are formed are subjected to an alternation of extremely high pressure and low pressure. The crack may become filled with lubricant, and as the ball rolls over the crack the oil is trapped and subjected to

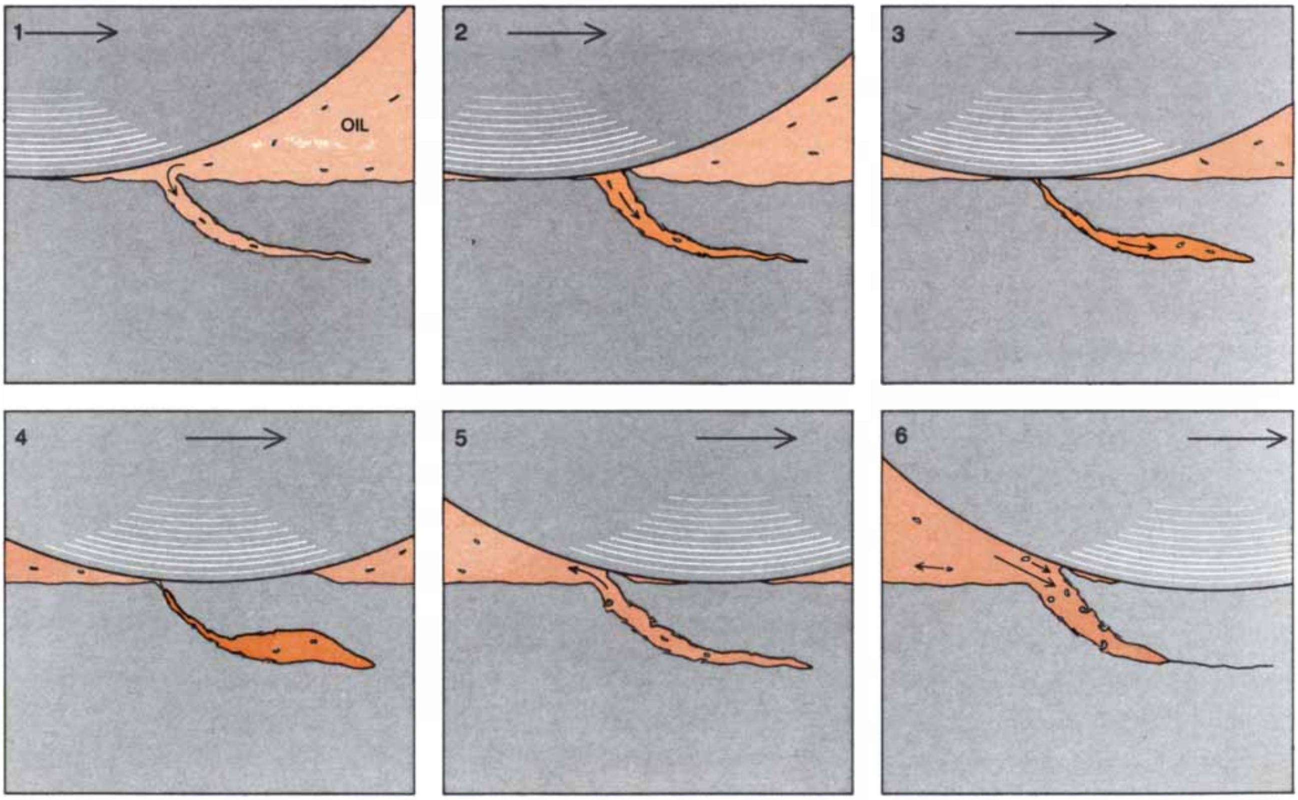


PIT ON SURFACE OF BALL BEARING, enlarged 25 diameters (*left*), was formed by spalling. The bottom surface of the crack from which the spalled material came is seen in greater detail at an enlargement of 120 diameters (*middle*). This surface is not the original fractured surface but rather a Beilby layer formed by the mechanical working of the crack. Spherical particles in the pit are shown enlarged 600 diameters (*right*). It is believed the spheres are held to the surface by a sticky polymer formed from components of the oil. These scanning electron micrographs were provided by Arthur W. Ruff, Jr., of the National Bureau of Standards.



SPHERICAL PARTICLES found on a Ferrogram slide are shown enlarged 3,500 diameters in this scanning electron micrograph. Smooth rubbing-wear particles are seen behind the larger sphere.

SPHERE IN CRACK appears to be in the process of formation. The surface of the sphere is covered with laminae that have not yet been smoothed out. Magnification is about 6,000 diameters.



FATIGUE CRACK in a ball-bearing race fills with lubricating oil. When a ball rolls over the crack, the trapped oil is compressed and subjects the interior of the crack to very high pressure. When the ball moves on, the oil is ejected. Repeated compression creates a Beilby layer inside the crack, from which flakes of metal break off. Minute spherical particles form in the cracks, perhaps as a result of a random rolling up of the small flakes generated in the crack or generated elsewhere and carried into the crack by the lubricant.

such high pressure that it may become a semisolid. As the ball moves on, the entrance to the crack opens up and the oil is ejected. The process is of course repeated an enormous number of times.

The motion of the surfaces of the crack and the stress produced by the pressurized lubricant can lead to the creation of a Beilby layer within the crack itself. Then thin flakes of metal quite similar to adhesive-wear flakes form in the crack. Scanning electron micrographs of cracks indicate that the spheres may be generated by a deformation process in which the flakes formed in the crack or generated elsewhere and carried into the crack by the lubricant are rolled into a spherical shape [*see bottom illustration on opposite page*]. The action can be likened to the process by which a schoolboy makes a spitball. The detailed mechanism whereby the random rolling of small flakes gives rise to spheres, however, is still unknown.

As lubricating oil circulates through a machine it picks up myriads of wear particles; indeed, so many wear particles are generated that a small sample of oil contains a complete record of how each wear surface is faring. Each particle in turn carries with it the history of the conditions under which it was formed.

When a machine is first put into service, it goes through a wearing-in process. During this time the absolute number of wear particles quickly reaches a peak and then decreases. Finally an equilibrium is reached, and the number and type of particles in the oil tend to remain fairly constant. The large majority of the particles in this equilibrium phase are rubbing-wear particles; the oil contains very few cutting-wear particles, large fatigue particles and spherical particles.

When new wear mechanisms arise, the particle distribution is modified. If the wear is severe, the amount of metal in the oil increases and the ratio of large particles to small ones goes up substantially, often by an order of magnitude or more. Simultaneously new types of particles other than rubbing-wear particles appear in quantity.

There is every expectation that with additional research it will be possible to reliably relate the wear particles in the oil with the part of the machine from which they come. Other research is directed to finding out more precisely how wear particles are formed and to determining the significance of various types of wear with respect to the future operation of the machine.

Computer Graphics in Architecture

A computer programmed to generate pictures or drawings can show a prospective building in various settings, enabling an observer to "walk" through the scene. It can also produce detailed plans

by Donald P. Greenberg

Drawings and models are the traditional means whereby an architect depicts to his client where a building might be put, how it might look and what the details of design and construction might be. The method is satisfactory but basically inflexible; if the architect or the client wants to see how the building would look in a different setting or orientation or with different details, new drawings and models must be made, at a considerable cost in time and money. Imagine, then, the benefits of a scheme that would quickly produce color pictures showing a prospective building in different places and perspectives, portraying, for example, how a new library on a college campus would look as one approached it on foot from various directions. Employing computer-generated graphic displays, my colleagues and I at the Cornell University College of Architecture, Art and Planning, in cooperation with the Visual Simulation Laboratory of the General Electric Company, have been working with a program that produces just such pictures. With further modifications, additional data and new plotting equipment the system would also be capable of rapidly printing detailed architectural drawings.

Our basic problem was one that artists have confronted since drawing began: How does one portray a three-dimensional scene on a two-dimensional surface? In Western cultures the early attempts to solve this problem entailed employing such standard cues of depth as the overlapping of surfaces and the fact that the size of a known object in a picture is inversely proportional to its distance from the observer. Artists have also developed such conventions as putting closer objects in the bottom foreground and making receding lines oblique.

Not until the Renaissance was the art of perspective drawing thoroughly investigated. The masters of the Florentine school produced many paintings in which perspective geometry was conveyed accurately and realistically. The usual technique was to have in the painting a single vanishing point, that is, a point where receding lines representing parallels seem to meet.

Today it is possible to derive this kind of visual logic mathematically and to instruct a computer to draw the perspective images. Consider an observer looking at a rectangular three-dimensional object [*see top illustration on page 101*]. The hypothetical lines from his eye to various points on the object constitute a bundle of view rays. The perspective drawing is constructed by determining the intersections of the view rays with an imaginary picture plane and connecting the appropriate points. The result is a two-dimensional image of the object projected on the picture plane.

A number of mathematical methods exist for computing a perspective image of an object composed of straight lines. They all require that a complete three-dimensional description of the object be stated in the form of the geometric coordinates of each vertex on the object. The frame of reference for the coordinates is a set of three axes (designated x, y and z) emanating from an arbitrary point and forming right angles with one another. The relation of the axes is what one would find in the lines emanating from a corner of a cube: a vertical line and two horizontal lines. Procedures for programming a computer to create the perspective image rely variously on solid geometry, vector equations and matrix transformations.

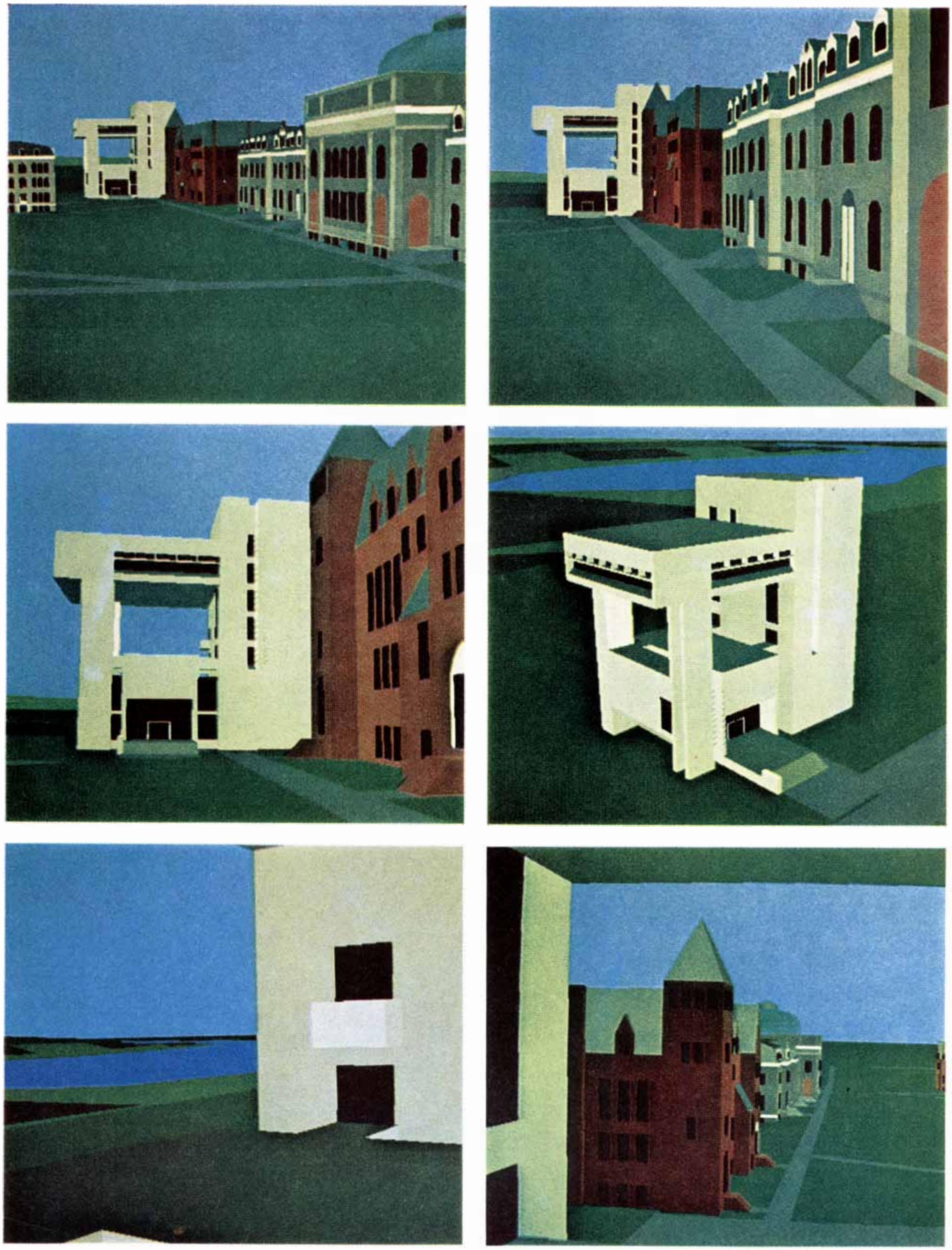
At present the commonest approach is to transform the mathematical quantities describing the observer, the object and the plane of the picture into a single coordinate system. The viewer is the origin, and the orthogonal axes are parallel to the picture plane and to the line of sight. By this means the intersection of the view rays with the picture plane can be computed readily by ratios derived from the geometry of similar triangles.

When all the lines of the object are drawn between their analogous points on the picture plane, the result is a "wire line" drawing [*see bottom illustration on page 101*]. This type of two-dimensional representation is confusing unless the "hidden lines" are removed. They are the lines the observer would not see in the three-dimensional object. Once the computer has been given the complex instructions on what it cannot "see," causing it to remove the hidden lines, more readable drawings can be generated.

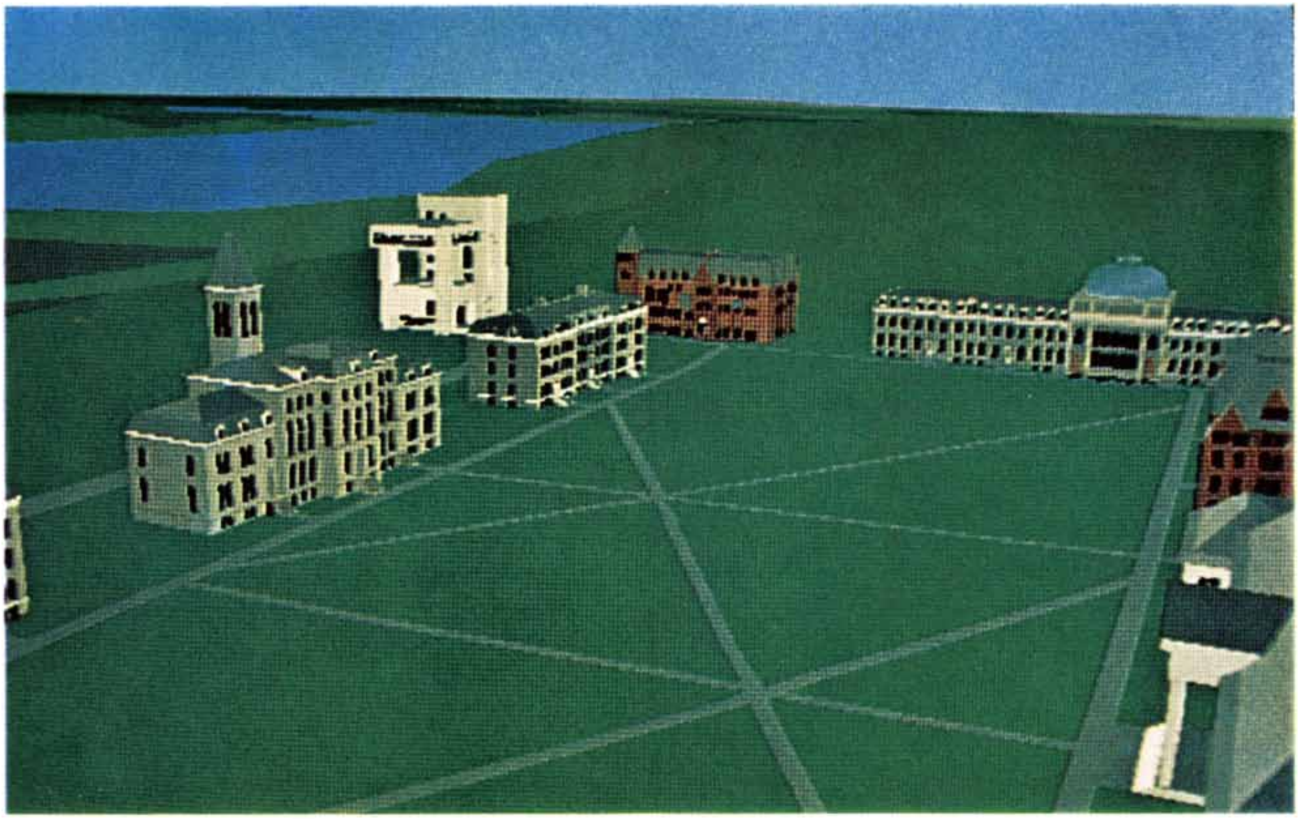
Drawings of this kind are usually obtained from the computer by means of a plotter in which the machine controls the motion of the pen or the paper or both. An electrostatic printer-plotter is another output device, which can produce both line drawings and halftone pictures similar to the ones yielded by photocopying machines. Still other devices for graphic output are the cathode ray tube and the television screen.

The most realistic graphic result is obtained with a color television set as the output device. This process was introduced by the Visual Simulation Laboratory of General Electric for simulation procedures designed to train astronauts for docking maneuvers in space. The abstract images were displayed 30 times per second, which is the rate of repetition of the picture on an ordinary television set and was rapid enough to simulate continuous movement.

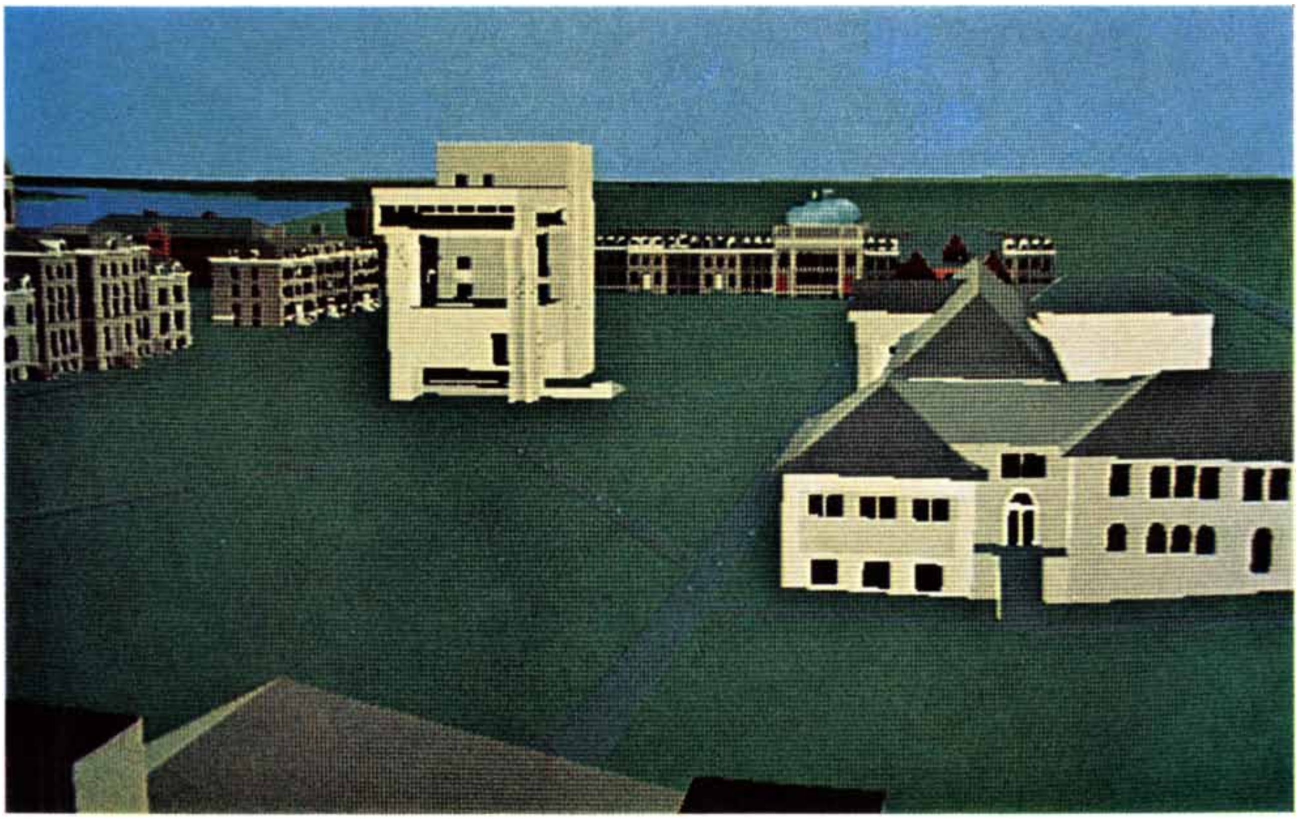
For architectural purposes such speed is not necessary. With less frequent displays highly complex and detailed im-



COMPUTER-GENERATED VIEWS of a building proposed for the Arts Quadrangle of Cornell University appear in photographs of a television display created by the computer. The building, a museum, is the squarish white structure in the background of the photograph at top left. In the first three photographs, reading from left to right, the observer is taken on a "walk" toward the museum. In the fourth photograph the museum is seen from a distant corner of the quadrangle. In the photograph at bottom left the observer is in the museum looking toward Cayuga Lake, and in the final photograph he is looking back along the route of the "walk."



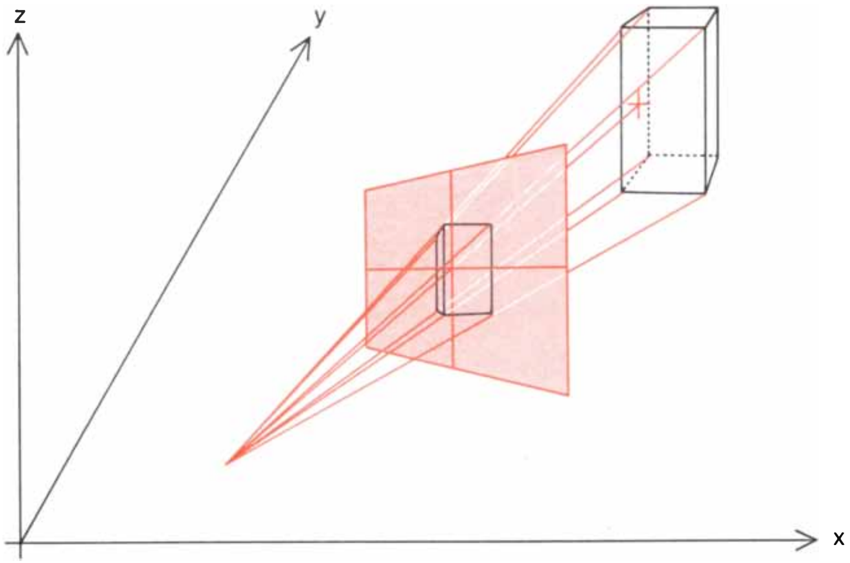
SETTING OF MUSEUM is shown correctly in this computer-generated view in which the observer is assumed to be on the roof of a building diagonally across the Arts Quadrangle. Route of walk in photographs on preceding page was along front of domed building.



MISLOCATED MUSEUM is placed by the computer in the middle of the Arts Quadrangle. The mislocation is deliberate and somewhat farfetched, but it demonstrates the ease with which different architectural schemes can be tested by means of computer displays.

ages can be constructed and filmed as a motion picture, thereby creating in the viewer a sense of walking toward a building or among a group of buildings. To demonstrate this capability our group produced two years ago, with the General Electric facility, a motion picture in color depicting the evolution over a period of time (future as well as past) of the Arts Quadrangle at Cornell. Most of the photographs accompanying this article are taken from the film.

In contrast to the kind of representation made by wire-line drawings the basic element for our system is the opaque plane. Lines are visible on the television screen only as the edges of two planes of different color. Since all the planes "painted" on the screen are opaque, the hidden-line problem becomes how to establish the proper sequence for displaying the planes. If as an observer looked at a building one plane obscured another plane or part of it, the obscured plane should be put on the screen first. In other words, the chronological ordering of the successive layers of the image is from background to foreground.



PERSPECTIVE DRAWING is the principle underlying the transformation of a three-dimensional scene into a two-dimensional picture by the computer. View rays (*color*) between an observer and a three-dimensional object intersect an imaginary picture plane. By connecting the appropriate points of intersection one obtains a representation of the three-dimensional object on the plane. The geometric framework whereby the computer locates an object in space is a system of coordinates arising from the orthogonal axes *x*, *y* and *z*.

A plane in such a situation can be described as a polygonal face that is either opaque or transparent to an observer, that is, he can see it or he cannot. For the computer the polygon is defined geometrically by its vertexes, and its color is identified by a particular number. Our convention was to put the vertexes describing a plane into the computer in a clockwise sequence when the plane was viewed from the opaque side and in a counterclockwise sequence when the plane was viewed from the transparent side. The computer treated the planes accordingly as opaque or transparent.

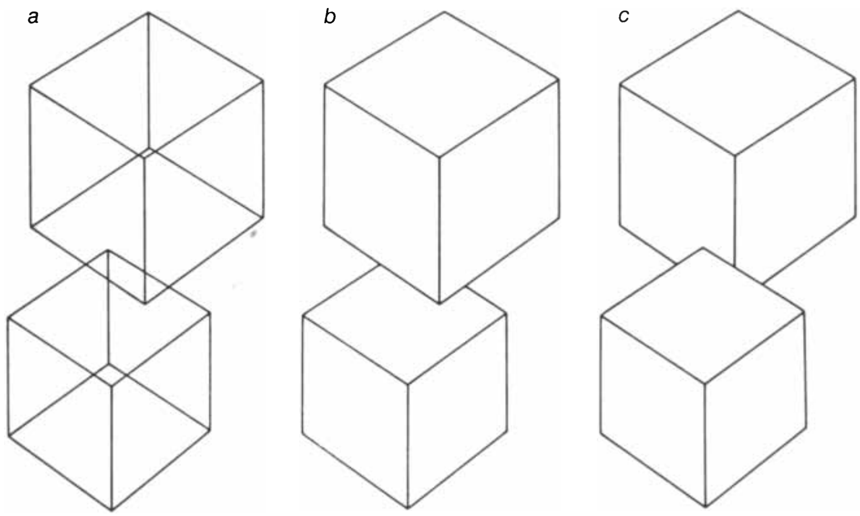
A plane or a face may also have subfaces that are contained entirely within the plane. Doors and windows are examples. The computations that determine when that plane should be portrayed in the sequence also serve for its subfaces, provided that each subface is painted over its parent plane.

The setting for a prospective building is likely to contain several objects, including existing buildings and parts of buildings and a certain amount of terrain. If in a complex setting it is impossible to determine the order in which the faces of an object should be painted, the object must be further subdivided into clusters by means of separating planes [*see top illustration on page 103*]. The planes establish a logic that serves the computer in determining the sequence of painting clusters. Within each cluster the sequence of painting the faces is unambiguous.

If the images are to be realistic, the computer must be given instructions for solving the hidden-plane problem. The mathematics for such a program are fairly complicated. In essence what the program must do is enable the computer to answer four categorical questions.

The first question is whether or not the plane is visible by itself. In answering the question the computer calculates the perpendicular to the surface of the plane. If the angle between the perpendicular and the direction of view is acute, the plane is visible [*see illustration on next page*]. If the angle is obtuse, the plane is not visible.

The second question is whether or not the plane is obscured by any other plane within the same cluster. If it is, the obscured plane is painted first, even if it is only partly obscured. On other occasions, as when one is looking toward a corner of a rectangular building, the two visible planes (the two visible sides of



HIDDEN-LINE PROBLEM must be dealt with by a computer program designed to yield clear representations of three-dimensional objects in displays of the "wire line" type. If all the lines of the object are displayed (*a*), the picture is confusing. The computer must be given instructions for removing the lines that could not be seen in a three-dimensional object or the result can be ambiguous. Removal of the obviously hidden lines in the figure at *a* could result in either *b* or *c* as the display unless the instructions were quite precise.

the building) do not overlap, and so the order in which they are displayed is inconsequential.

The distinctive property of a cluster is that all the faces in it can be properly ordered for display regardless of the position of the observer. This property can best be seen by referring again to the illustration depicting separating planes and clusters [*top of opposite page*]. The object in the illustration is a schoolhouse that has been subdivided into three clusters by means of two separating planes.

It is also possible, of course, for the relation between planes to change as the observer moves. Imagine a situation in which Plane *A* obscures part of Plane *B* when the observer is in one position, whereas Plane *B* obscures part of Plane *A* when the observer moves to a different place. In this case it is impossible to order the sequence of display on the basis of nothing more than the geometry of the object, since the sequence also depends on the position of the observer. Here a separating plane is devised to eliminate the ambiguity.

I should emphasize that separating planes and clusters are concepts, not real objects; their purpose is to make it possible for a computer to display plane faces in a sequence that yields a realistic picture. Determining the location of the separating planes for a given structure requires a fair amount of skill and practice. Once it has been done, however, the faces in each cluster can be ordered without further concern for the position of the observer. The computational process is rather long, but it need be done only once. The resulting ordered list of faces is stored in the computer's memory.

Now the third question arises: Is the cluster obscured by any other cluster within the same object? The order in which the clusters are displayed depends on the position of the observer. In the illustration of the schoolhouse the observer's position is on the near side of Separating Plane *1* and above Separating Plane *2*. It is evident that from his perspective Cluster *B* will obscure parts of Cluster *A* and Cluster *C* and Cluster *A* will obscure part of Cluster *C*. Since the order of display is from background to foreground, the sequence of painting the clusters for the schoolhouse is *C*, *A* and *B*.

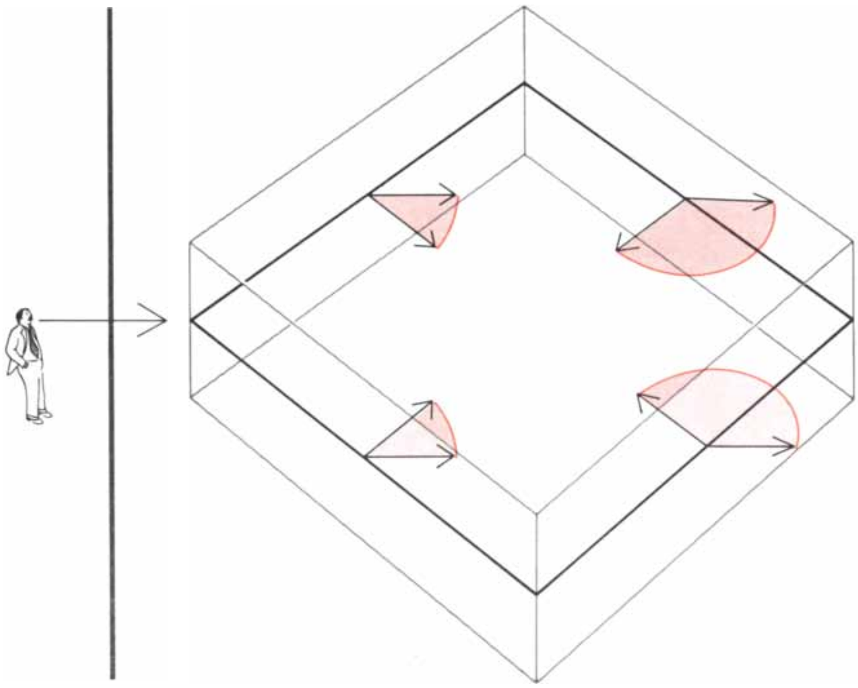
The final question is whether or not the object is obscured by any other object. It therefore concerns the order in which objects are displayed. Usually objects are separate (two buildings, for example), so that one can deal with them by means of range points. Each object is arbitrarily given a range point. The order in which the objects are displayed is then determined by their distance from the observer. The objects farthest away will be painted first, and the closest ones will be painted last.

Once the perspective transformations and the sequence for displaying images have been established for a particular scene, the picture can be formed on the television screen. The computer can control the generation of the picture in such a way that each dot or raster on the screen will be given the appropriate color. The accompanying series of photographs [*page 104*] shows how an image is built up. The time needed for the computer to generate the finished image is about two seconds.

A major advantage of this approach is that it can compute the portion of the priority list that is dependent on the geometry of the object and independent of the position of the observer. These are the calculations performed in solving the relations among planes within clusters. As a result a substantial part of the computational work is done before the picture is generated.

Recent studies by workers at General Electric and by others, notably Ivan E. Sutherland and his colleagues at the University of Utah, have brought advances in computer displays. One improvement has increased the efficiency and versatility with which halftone images can be generated. Another has created smooth shading techniques, which eliminate hard edges so that rounded polygonal surfaces can be made to appear rounded in the display. Programs for depicting illumination, shading and shadowing, transparency and the reflection of light have been developed. Graphic devices giving higher resolution and therefore clearer pictures are being tested.

For architecture the rapid improvement in the technology of graphic display promises a number of benefits beyond being able to show how a building will look in its setting and to simulate a walk through the scene by an observer. For example, if information is stored in the computer's memory on specific features of the design, on geometric descriptions of architectural details, on standard components such as windows and on estimating costs, it will be easy for the machine to generate both working drawings and cost surveys. Moreover, given programs for spatial evaluation or for structural and mechanical analysis, the machine will be able to



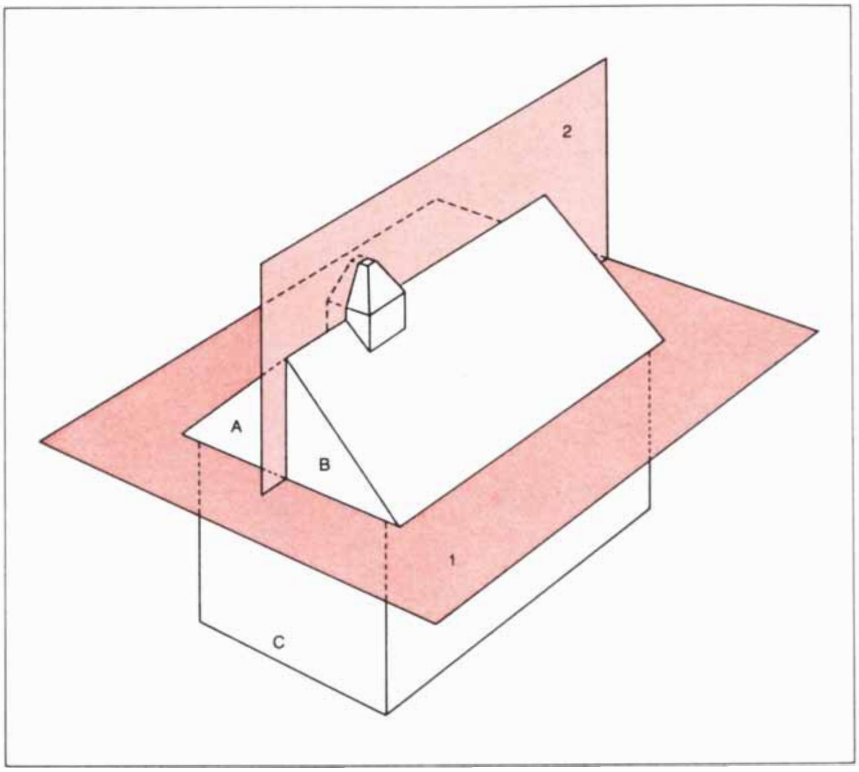
VISIBLE SURFACES of a three-dimensional object must be established for a computer if the machine is to "paint" a realistic picture of a building on the television screen that serves as the output device. The program that enables the computer to ascertain what faces would be obscured is based on the angle (*color*) between the perpendicular to each face and a line parallel to the observer's line of sight. If the angle is acute, the face is visible from the position of the observer, whereas an angle that is obtuse signifies an invisible face.

serve as an interacting partner in the design process.

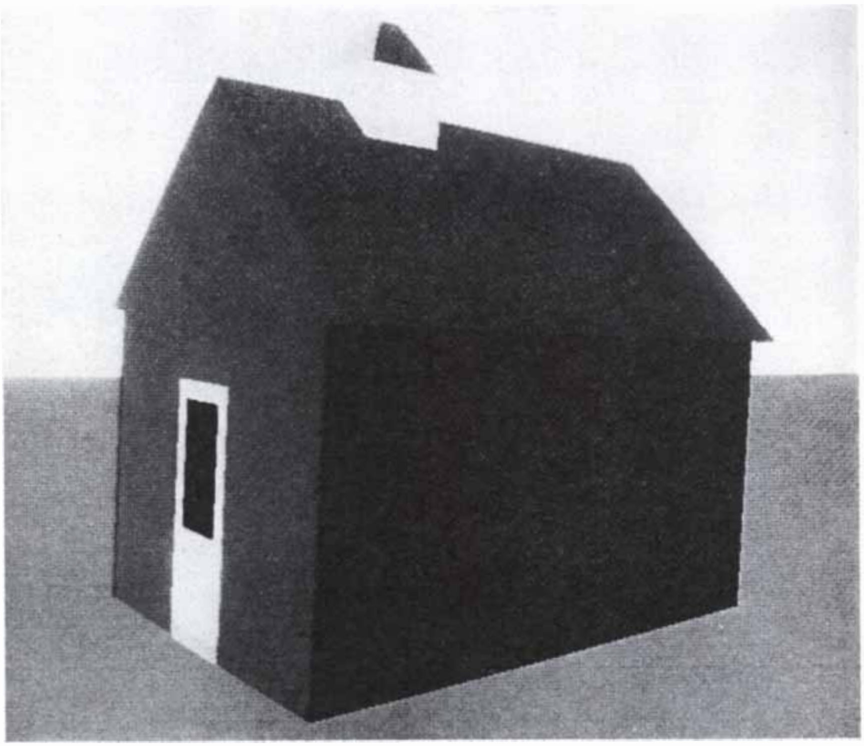
The typical stages of an architectural design process are schematic design, in which floor plans are roughly sketched; preliminary design, in which the elements of the plan are assembled into a representation of the completed building, and final design, which includes working drawings and models. In the first two stages interactive computer processes are already available. Nicholas P. Negroponte of the Massachusetts Institute of Technology has been working on sketch-recognition techniques, which enable the designer to "doodle" with plans and have his intentions interpreted correctly by the machine. The firm of Perry, Dean and Stewart in Boston employs computer graphic displays in an interactive way to determine the relation the various parts of a building should have to one another. The Chicago office of Skidmore, Owings & Merrill has a computer arrangement that prints out representations of the city's downtown Loop area; the firm can utilize the pictures in preliminary design studies.

Computer-based drafting and graphic displays of other types have proved to be accurate, economical and rapidly producible. As a result their contributions to architecture have been increasing. An example is a machine-based drafting system employed by Saphier, Lerner, Schindler, Environetics, Inc., of New York to generate all the interior-design drawings for the Sears Tower in Chicago. Four separate plans (for construction, for the reflected ceilings, for finishing and for furniture, telephone and electric arrangements) were prepared for each of the 110 stories [*see illustration on page 106*]. The large amount of repetition from one floor to the next made this computer operation particularly advantageous. The most complicated plan, which might ordinarily take one man-week of drafting time, was done by the machine in 45 minutes. The firm has found the system so effective that it is now producing all its working drawings in this way.

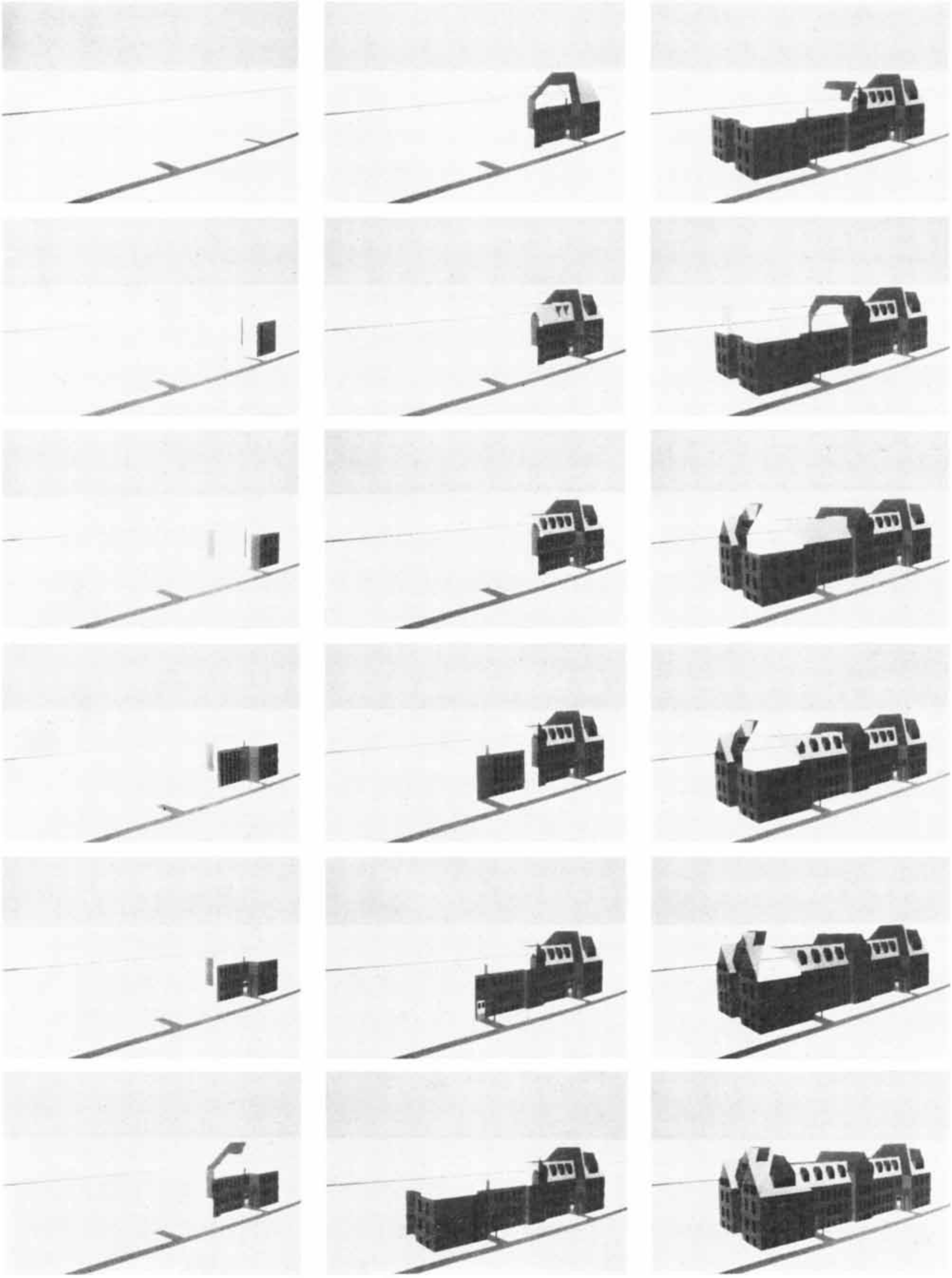
A system that produces computer-generated images in color offers three major advantages over the standard architectural drawings and models. They are flexibility in changing the design, the ability to simulate motion and the opportunity to change and experiment with color. These advantages are evident in the accompanying photographs [*page 99*] relating to the construction of the Herbert F. Johnson Art Museum at Cornell. To cause the computer to gener-



SEPARATING PLANES AND CLUSTERS are concepts employed to enable a computer to generate a realistic picture when the visibility of a face of an object depends on the position of the observer. In this example the observer is at a point in the center foreground. For programming the computer the structure, which is a one-room schoolhouse, is divided conceptually by two separating planes (*1 and 2*) into three clusters (*A, B and C*). The distinction of a cluster is that all the faces in it can be ordered properly without reference to the position of the observer. The order in which the computer displays clusters, however, depends on where the observer is. Here the order of clusters would be *C, A* and *B*.



PORTRAYAL OF SCHOOLHOUSE by the computer resulted in this picture. Same program could produce a suitably changed picture for a different location of the observer.



COMPUTER PORTRAYAL OF BUILDING is shown step by step as it would be displayed on a television screen. The building is Lincoln Hall, which is part of the Arts Quadrangle at Cornell. Once the program for such a display has been put into the computer, the picture can be generated quickly. The buildup of the display shown here took about two seconds. The computer also controls the generation of the picture in such a way that each dot is given the appropriate color from an array of 64 color possibilities.

ate the images, each building and object in the environment (the museum is on the periphery of the Arts Quadrangle) was mathematically modeled in its own three-dimensional coordinate system.

If for the coordinate system of each object the coordinates of the origin and orientation of the axes are known, then by means of the appropriate translations and rotations the data can be combined into one general system representing the environment of the buildings and objects. The advantage of this hierarchical structure of coordinate data is that one can then vary the site of each building easily, thereby testing with the computer a variety of different solutions to problems of design [*see illustrations on page 100*].

By similar means it is possible to specify a moving coordinate system and have an object move through the environment. We did this in portraying a streetcar that once ran on the Cornell campus. It is also possible to change the dimensions or the scale of a building by altering the numerical input relating to the geometry of the building, although this procedure is somewhat more complicated.

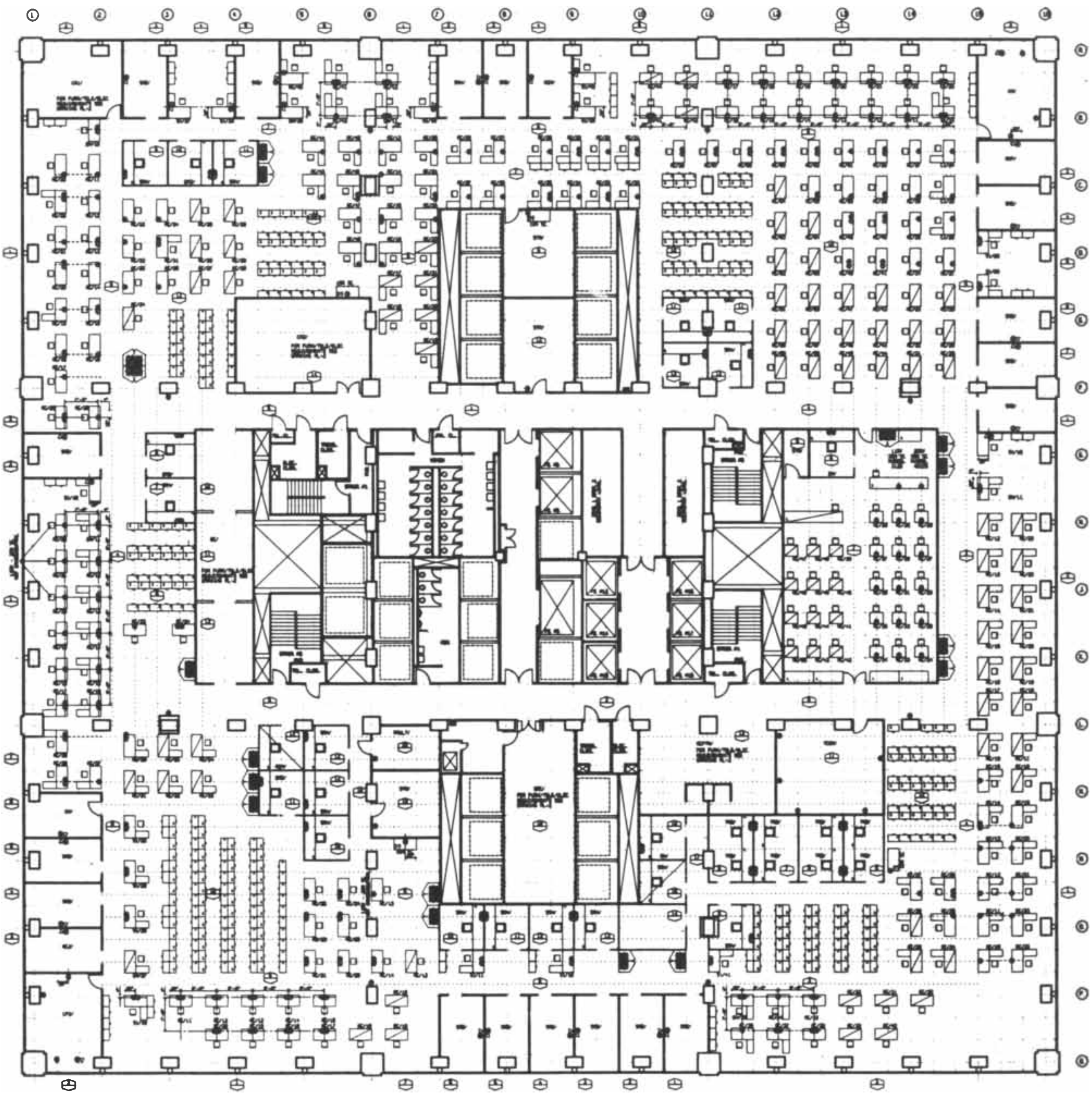
For architects the ability to simulate motion is highly useful. One of the principal concerns of architectural design is space: the internal spaces of a building and the external space of the building and its setting. One does not react to space from a static position, as one might view a painting. To obtain a deeper understanding of architectural space it is necessary to move through the space, experiencing new views and discovering the sequence of complex spatial relations. By filming a sequence of computer-generated images one can obtain a realistic simulation of a walk through a space, even if the space is in or around a building that has not yet been built.

Once the environment has been modeled mathematically, motion can be simulated by describing changes in the position of the observer and the direction in which he is looking. If one supplies the computer with this information for the starting and finishing points of the walk and with the total number of images desired from start to finish, the computer can do the rest, calculating the coordinates of each position occupied by the imaginary stroller and the direction of view from it. This procedure will yield a walk of constant velocity. To achieve smooth transitions instead of a jerky motion, particularly at the starting and finishing points, one can supply the computer with acceleration functions that increase or decrease the simulated velocity as desired.

Since each color can be identified mathematically, the system enables the architect to experiment with color. All faces of the resulting image with the same number will be displayed in the same color. The intensity, hue and saturation of each color can be varied by controlling the components of red, green and blue light in the display. A six-bit code can specify up to 64 different colors for any particular scene by assigning two digits each to the three color components.

A further refinement that is available but was not included in our project is to have the machine compute the particular colors. Suppose the exterior of a proposed building is to consist of one textured material, say a gray concrete. A range of gray shades can be assigned the numbers 1 through 10, which is a sufficiently large range to provide for a considerable variation in the display. For generating pictures a specific shade will be given each plane in the image according to the angle between the sun and the perpendicular to the plane. Therefore all faces of the building can be shaded automatically by the machine.

We foresee that as the availability of computer graphic systems continues to increase and their cost continues to decrease, the process will be employed increasingly in architecture both as a medium of presentation and as a means for generating design drawings. Even more important is the potential of computer-generated images as a tool for design and analysis. The ability to test an environment before it is created is of great benefit to architecture and planning. The techniques for doing so are only now beginning to be explored.



ARCHITECTURAL DRAWING of the furniture, telephone and electric plan for a floor of the Sears Tower in Chicago was generated by a computer under a program developed by Saphier, Lerner, Schindler, Environetics, Inc. Drawings were made for four separate plans for each of the building's 110 stories. Large amount of repetition from floor to floor made such drawings particularly useful.

What would you do with a Statistics and Number-Crunching Computer that starts at $7,100,* has 16K Hardwired Basic Language and 28 Major Peripherals?

Plenty!

The new Wang System 2200 is a System. It gives you the raw power and the peripherals you must have for a wide range of problem solving. For under $7,200 you get a CPU with 16K bytes of BASIC language instructions hardwired into the electronics . . . plus a 4K operating memory. You also get a big 16 lines (of 64 characters each) CRT display, a console mag tape drive and your choice of either alpha or BASIC Keyword keyboards.

Some Words About Language: The hardwired MOS ROM language in your System 2200 finally ends your dollar tradeoffs . . . economy systems that are costly to program or very expensive systems that are relatively easy to program. Many, if not most, of your people already know BASIC. They'll be solving problems the day your system is delivered (and, we can deliver in about two weeks). Most of your budget will go into problem solving; not system support.

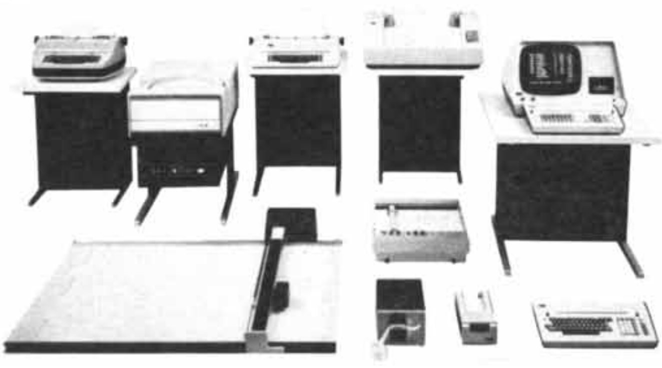
Try To Out-Grow It: Main memory is field expandable in 4K increments (at $1,500 per 4K). Up to 32K. You can choose from three kinds (and 7 price ranges) of printers . . . one even has a stepping motor for very precise 4-quadrant incremental plotting. Speaking of plots, we have a new, very large flatbed (31" × 42") for only $7,500 or a smaller one if you plot small. Both print alphanumerics and plot under full program control. Been appalled lately by disk prices? Starting at just $4,000, we offer you our new "floppy" disk in single, double and triple disk configurations (.25, .50 and .75 MB's). For big disk power, you can have 1, 2 or 5 megabyte fixed/removable disk systems. All peripherals, including punched or mark sense hopper card readers, paper tape readers and on- line BCD or ASCII controllers are easily added-on in the field so your System 2200 will grow with your needs.

The Wise Terminal: If you are now or may soon be getting into terminals, we have several new products that will instantly upgrade your System 2200 for telecommunications with any other System 2200 or a mainframe computer. And, you still have a powerful stand-alone system. Another approach, of course, is to justify it as a powerful terminal and get a "free" stand-alone computer. Wise?

We Do A Lot For You: System 2200 is backed by over 250 factory-trained Wang Service Technicians in 105 U.S. cities. Naturally, we guarantee or warranty everything you buy from us. If you want, there are free programming/operating schools here in Tewksbury, Massachusetts, almost every week. We have a growing program library on a wide range of statistics and math/science applications. Our user group (with the unlikely name of "SWAP") could help you cut programming costs even further. We do a lot for you.

* All prices U.S. List. If you're the entrepreneur type, we've just announced a new 7-module Basic Accounting System software package for the business end of your business like payroll, invoicing, inventory, receivables and some really fancy management reports.

WANG



Even if you call the Wang System 2200 a small system . . . you have to admit it's a big idea.

☐ Please send literature

☐ Please have your representative phone me:__________

Name____________________Title__________

Company______________________________

Address______________________________

City_________________________________

State_________________________Zip________

SA-5

Wang Laboratories, Inc.,
836 North St., Tewksbury, Ma. 01876, Tel. (617) 851-4111

DEUTERIUM IN THE UNIVERSE

All the heavy hydrogen in space may have been made in the first 15 minutes after the "big bang." Observations leading to estimates of its abundance thus provide evidence on conditions at that time

by Jay M. Pasachoff and William A. Fowler

The calculations of modern cosmology lead to the conclusion that the universe originated between 10 and 20 billion years ago. The particular cosmology that seems to have the greatest weight of observational evidence behind it is the "big bang" theory, which posits that all the matter in the universe was compressed into a superdense kernel that exploded and has been expanding ever since. Within the past few years evolving techniques of observation have made it possible to look backward in space and time to what appear to be the first few minutes after the big bang, and to learn about the physical state and early evolution of the universe.

One item of observational evidence in favor of the big-bang theory was the discovery that the universe is permeated uniformly by radiation corresponding to what would be radiated by a theoretically perfect "black body" with a temperature of three degrees Kelvin (degrees Celsius above absolute zero). This radiation is believed to be a remnant of the original big bang. The newest evidence has come to the fore within the past two years. It is the detection in interstellar space of atoms that may have been formed within the first 15 minutes after the big bang. These atoms are the atoms of deuterium, commonly known as heavy hydrogen.

The big-bang theory stemmed from the discovery by Edwin P. Hubble 50 years ago that all other galaxies appear to be receding from ours, and that the most distant galaxies seem to be receding the fastest. The big-bang theory accounts for Hubble's observations by the expansion of the universe, as does the steady-state cosmology advanced by Hermann Bondi and Thomas Gold and independently by Fred Hoyle. The steady-state theory, however, postulates that the universe has always looked the same as it does today, and it requires that hydrogen be continuously created to make up for the decreasing density of the universe caused by its expansion. Most astronomers believe this explanation has been ruled out by the discovery of the three-degree background radiation. A variant big-bang cosmology posits that the universe will stop expanding and then contract. The recent measurements of interstellar deuterium have provided a method for determining whether that will happen or whether the universe will simply go on expanding forever.

The "standard" big-bang theory states that in the first 100 seconds after the initial explosion of the universe only protons, neutrons, electrons, positrons and various types of neutrinos existed and that each particle was independent. As the exploding universe cooled, the protons and neutrons began to combine. Since the nucleus of an atom of "ordinary" hydrogen is simply a proton, the combinations of protons and neutrons were the nuclei of the heavier isotopes of hydrogen and of the other atoms. The simplest compound nucleus is the deuteron, composed of one neutron and one proton, which with a single electron forms an atom of deuterium. The nucleus of the third isotope of hydrogen, tritium, consists of one proton and two neutrons. The combination of two protons and one neutron forms the nucleus of an atom having a quota of two electrons; it is the light isotope of helium. The nucleus of the common isotope of helium has two protons and two neutrons. The nuclei of heavier elements are built up in the same way.

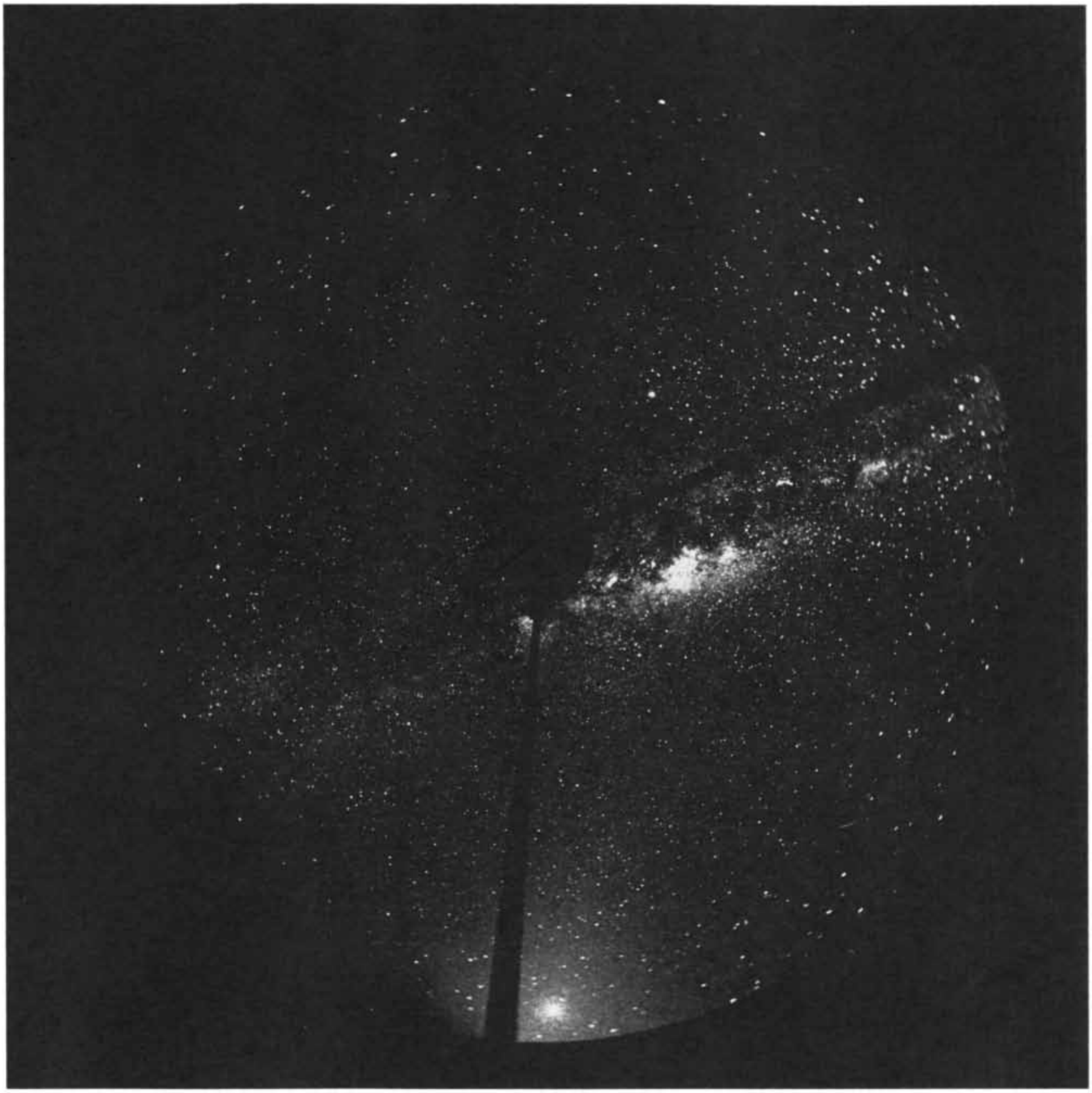
In 1948 Ralph A. Alpher, Hans Bethe and George Gamow first calculated the way in which the nuclei of the lighter elements could have been synthesized in the big bang. In 1957 one of us (Fowler), together with Geoffrey Burbidge, E. Margaret Burbidge and Hoyle, laid out a blueprint showing how both the lighter and the heavier elements could have been built up in the interior of stars. Seven years later Hoyle and Robert J. Tayler calculated that the large amount of helium observed throughout the universe (between 20 and 30 percent of the total mass) could not have been manufactured in ordinary stars. The nuclei require a temperature of 10 billion degrees K. for their synthesis, a temperature that could have been attained only in the big bang, in the explosions of supermassive stars ("little bangs") or in supernovas, the explosions of ordinary stars.

In 1966, after the discovery of the three-degree background radiation, P. J. E. Peebles tried to calculate what the relative abundance of deuterium and helium would have been if those two elements had been formed in the big bang. At the same time Robert V. Wagoner joined Fowler and Hoyle to check whether or not the new data sufficiently changed the outcome of previous calculations enough to allow heavy elements to have been synthesized in the big bang. They found that it was still not possible. Could these elements have been formed in events other than the big bang? Let us deal first with their synthesis in the big bang and then discuss the possible exceptions.

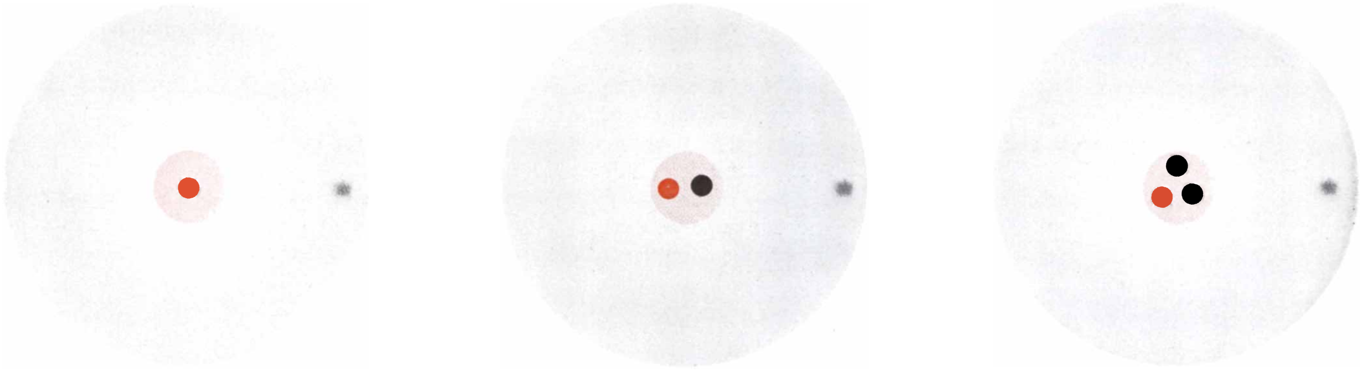
The calculations of Wagoner, Fowler and Hoyle show that there is only one isotope left over that is unique to the big bang. It is deuterium. Although helium would also have been made in the big bang, additional helium has been synthesized since then in stars and cannot now be distinguished from the original helium. Deuterium, however, is only depleted by the processes that go on inside ordinary stars; it is "cooked" into heavier elements by thermonuclear reactions. Therefore whatever deuterium we find now may well have persisted since the origin of the universe. The deuterium would have been formed in the first 1,000 seconds after the big bang, when the rapidly expanding universe had cooled off just enough to allow protons and neutrons to combine. In some sense, then, deuterium may actually allow us to look backward 10 to 20 billion years in time.

In order to use deuterium as a key to gaining a fuller understanding of the big bang, it is necessary to determine how abundant deuterium is with respect to ordinary hydrogen. Ideally we should like the cosmic abundance ratio, if it exists as a meaningful constant, to be uncomplicated by processes inside stars or by the chemical interactions of elements and compounds on planets or on grains of ice or dust in space. The ratio is best measured in the gas between the stars.

The density of the interstellar gas is very low: about one atom per cubic centimeter. The space between the stars is so vast, however, that the total amount of gas is very large. It is well known that almost all this gas is hydrogen. The problem is to find out if any of the hydrogen is deuterium and if so how much.



WIDE-FIELD PHOTOGRAPH OF THE MILKY WAY, made by Wolfhard Schlosser of the Ruhr-Universität at Bochum in Germany, shows the region of the center of the galaxy in the direction of which the authors and their colleagues conducted their search at radio wavelengths for absorption from interstellar deuterium. Galactic center is in the constellation Sagittarius, just above knot of stars to right of center. Round dark object is secondary mirror of astronomical camera. Three linear forms are the mirror supports.



ISOTOPES OF HYDROGEN AND HELIUM differ in the number of their nuclear particles. The lightest isotope is ordinary hydrogen (*left*), which has a single proton in its nucleus (*dark color*) and a single electron (*dark gray*). Deuterium has a proton and a

How can deuterium be observed? In an atom of ordinary hydrogen the single proton and single electron act as though they are spinning like tiny tops. The spin of the electron can be either in the same direction as the spin of the proton or in the opposite direction. If the spins are in the same direction, the atom is in a higher energy state, and the electron spin can flip over so that it is spinning in the opposite direction with respect to the proton. This spin-flip emits a small amount of energy at the radio wavelength of 21 centimeters. Gaseous hydrogen can also absorb energy at this wavelength as well as emit it. The atoms that absorb the energy are thus put into the higher energy state. Although the spin-flip for any one atom is rare, there are enough atoms in the galaxy for the radiation at 21 centimeters to be quite strong.

Deuterium also has a single electron, and the radiation from its spin-flip is at a wavelength of 92 centimeters. Deuterium, however, is much less abundant than ordinary hydrogen. In the water of the earth's oceans there are 6,600 atoms of ordinary hydrogen for every atom of deuterium. The abundance of deuterium in the solid body of the earth could be quite different. Thus the value for the oceans does not necessarily reflect the abundance ratio of deuterium to ordinary hydrogen on the earth, much less the ratio in interstellar space.

Soon after the spectral line of hydrogen at 21 centimeters was discovered, several astronomers began to search for the line of deuterium in interstellar space. The surveys culminated 10 years ago when Sander Weinreb of the National Radio Astronomy Observatory (NRAO) at Green Bank, W.Va., turned the observatory's 85-foot radio telescope to the constellation Cassiopeia to look for the 92-centimeter line in the spectrum of the strong radio source Cassiopeia A. Although Weinreb observed for weeks, he was totally unable to detect deuterium. He stated on the basis of his negative observations that the ratio of deuterium to ordinary hydrogen in that direction appeared to be less than 1 : 13,000, since any larger amount would have been detected. That upper limit is half the ratio of deuterium to ordinary hydrogen on the earth.

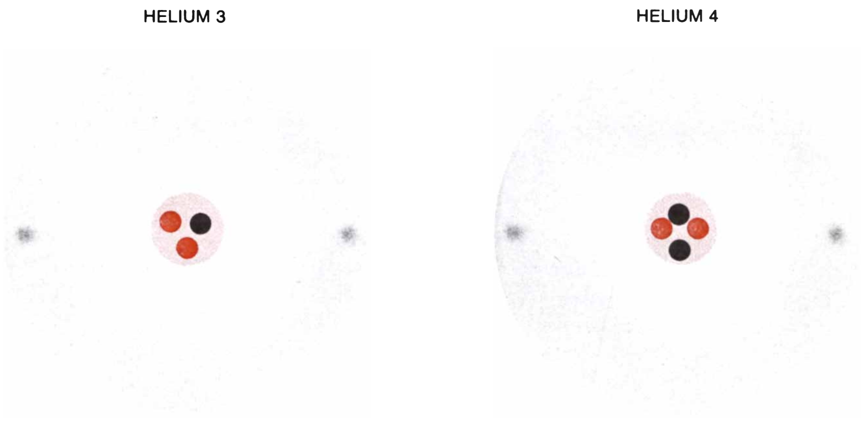
Radio astronomy has continued its rapid development since the time of Weinreb's survey, and within the past few years many complex molecules have been detected in the clouds of gas and dust in interstellar space. Most of the molecules have been found at relatively short wavelengths, from about 21 centimeters down to two millimeters. Few spectral lines have been discovered at wavelengths longer than a meter, although the failure has not been for lack of trying. In 1969 Carl A. Gottlieb, Dale F. Dickinson and one of us (Pasachoff) used the 150-foot radio telescope of the Air Force Cambridge Research Laboratories to search at meter wavelengths for various molecules, some of which included deuterium as a constituent. Two such molecules were the hydroxyl radical (OH), with deuterium in place of ordinary hydrogen, and water, with deuterium in place of one of the two ordinary hydrogens. The search was later continued at the NRAO in collaboration with David Buhl, Patrick Palmer, Lewis E. Snyder and Ben Zuckerman. No molecules were detected.

The chance of finding the faint spectral line of deuterium, or for that matter any other spectral line at such wavelengths, seemed very small. Nevertheless, in conversations in 1970 the authors of this article felt that another attempt should be made. The project would call for many weeks of observing, but it seemed that the 130-foot radio telescope of the Owens Valley Radio Observatory of the California Institute of Technology might be available. Diego A. Cesarsky and Alan T. Moffet of the staff at Owens Valley joined one of us (Pasachoff) in making the observations. In the following discussion of these observations "we" refers to Cesarsky, Moffet and Pasachoff and students from Williams College and Cal Tech who worked with us.

We decided to observe for two weeks at a time because we wanted to analyze the results as we went along. Many problems could have interfered with the investigation. For example, the 92-centimeter line lies in a region of military air-to-ground communications. Although the aircraft transmissions were sometimes bothersome, we were able to remove those time periods from our data.

The first two weeks of observing were in March, 1972. The main direction in which we chose to observe was toward the center of our galaxy. The galactic center, however, is in the southern sky, and at the latitude of the Owens Valley it is above the horizon for only about six hours a day. Since radio telescopes can operate in broad daylight, we were able to observe the galactic center whenever it was above the horizon, to observe in the direction of the Great Nebula in Orion much of the rest of the time and to observe in the direction of Cassiopeia A during any intervening hours. The data from this first observing run looked good, showing that the telescope and the receiving system were working satisfactorily. Futhermore, there was actually a suspicion of a faint absorption line at 92 centimeters.

That summer we had three more observing runs of two weeks each, extend-

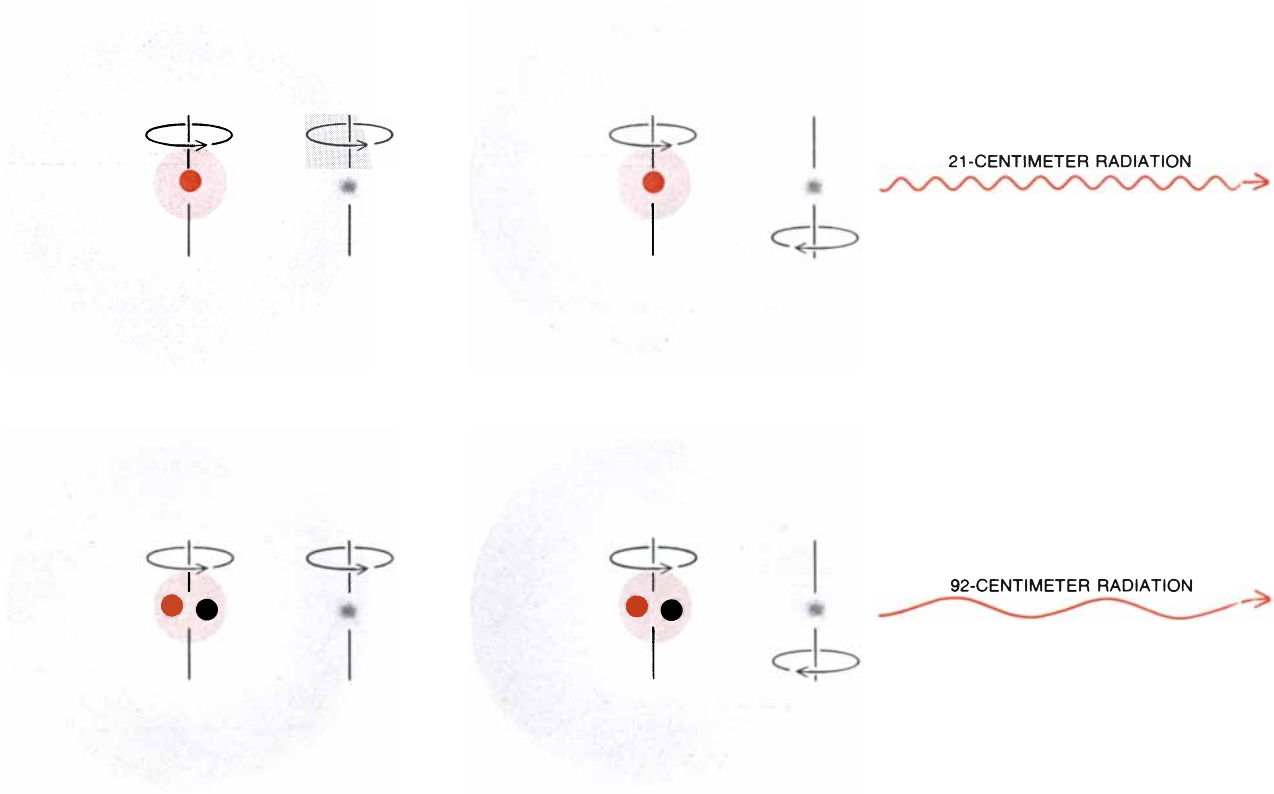


neutron (*black*), tritium has a proton and two neutrons. Helium 3 has two protons and one neutron in its nucleus and two electrons. Helium 4 has two protons and two neutrons.

ing the results of the first run. We concentrated our efforts in the direction of the galactic center. There is more hydrogen in front of that source than in any other part of the sky. We assumed as a working hypothesis that the abundance ratio of deuterium to ordinary hydrogen is constant throughout the universe, so that presumably the greatest total amount of deuterium would lie in the same direction as the greatest total amount of ordinary hydrogen. The galactic center is a strong source of radio waves, and the gas in the 40,000 light-years between it and the earth absorbs some of the radiation at the wavelengths of 21 and 92 centimeters.

Our analysis of the data by computer during that summer and fall strengthened our belief that we were actually observing the absorption line of deuterium. In the summer of 1973 five more weeks of observing by Cesarsky and Pasachoff yielded results that were compatible with the findings in 1972. We attempted to observe over a wider range of wavelengths, but it seemed as though we were now encountering too much outside interference in the wider band. In the narrower range of wavelengths observed in common during both years, the largest absorption fluctuation again corresponded to the spectral line of deuterium at 92 centimeters.

The line is still only barely visible in the data, which show expected random fluctuations [*see top illustration on page 113*]. Adding the results of both years' observations together has slightly improved the ratio of the signal to the noise and the line seems more than three times as strong as the fluctuations. Even if the "line" turns out to be only a particularly large random fluctuation in the data, the magnitude of the surrounding noise places an important upper limit on how strong the absorption could be. From our data that upper limit is one part in 3,000. If the feature we observed is the deuterium line, then the abundance ratio of deuterium to ordinary



SPIN OF THE ELECTRON FLIPS on rare occasions in an atom of ordinary hydrogen or deuterium, with the emission of radiation characteristic of the atom. In the schematic atom of ordinary hydrogen at top left the nucleus and the electron are spinning in the same direction. In the second drawing at top right the electron has flipped so that it is spinning in the opposite direction, with the emission of radiation at a wavelength of 21 centimeters. The same kind of spin-flip in the deuterium atom (*third and fourth drawings*) gives rise to 92-centimeter radiation. When the atoms absorb radiation of same wavelength, the spin of their electron flips back.

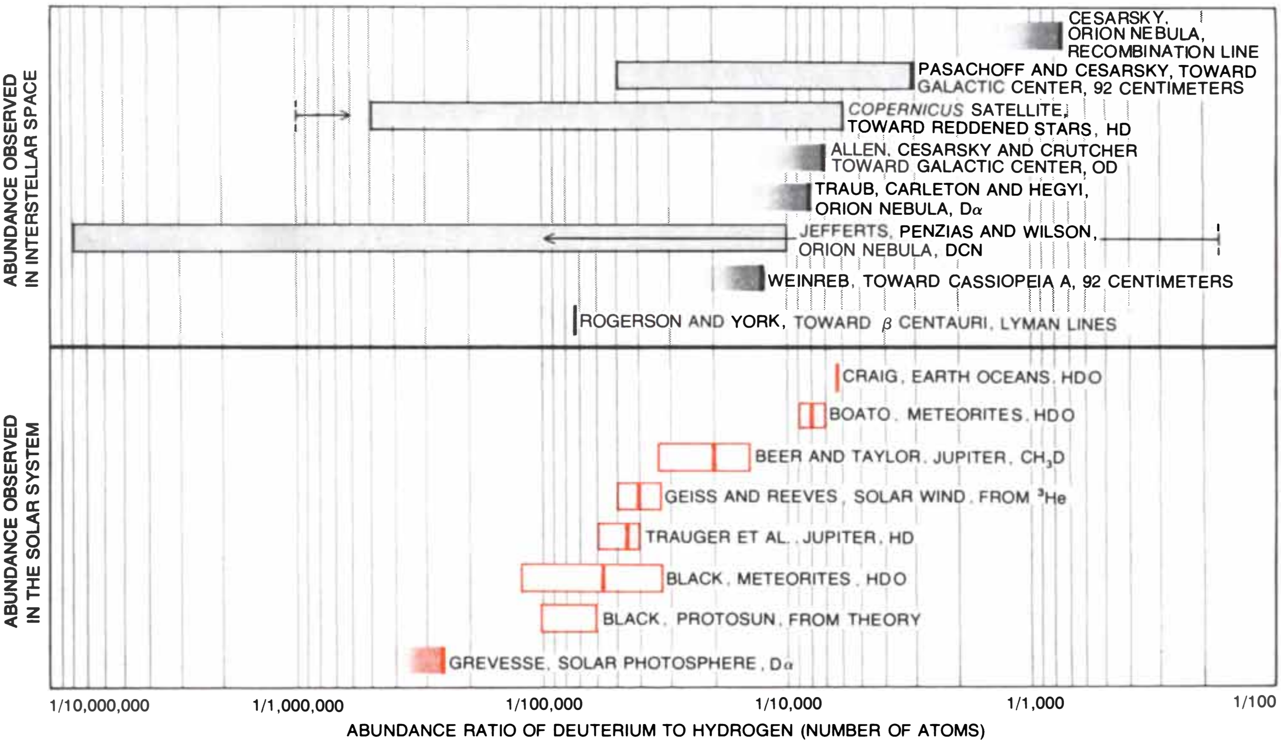
hydrogen would be between 1 : 3,000 and 1 : 50,000.

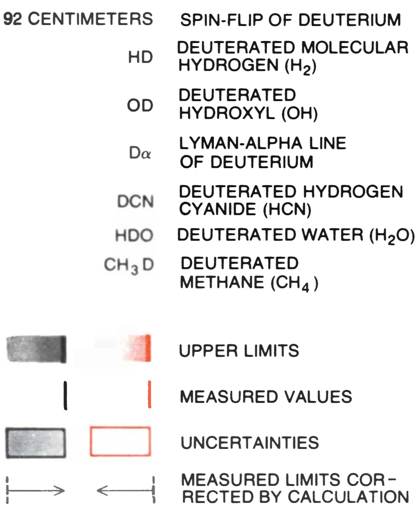
Curiously the major part of the uncertainty in that range arises not because the absorption line of deuterium is so weak but because the radiation from ordinary hydrogen with which it is compared is so strong. In the distance between the galactic center and the earth there is so much hydrogen that the radio signal is saturated. The result is that one cannot tell exactly how much hydrogen there is in this space, and one can only determine a lower limit. The next step in improving the determination of the abundance ratio of deuterium to ordinary hydrogen is not only confirming the detection of the absorption feature at 92 centimeters but also improving the measurements of the strength of the hydrogen radiation at 21 centimeters.

The gas seems to have the small velocity of 3.7 kilometers per second in our direction. The spectral lines of other gases observed in the direction of the galactic center, notably the hydroxyl radical and formaldehyde (H_2CO), are also Doppler-shifted by an amount that corresponds to the same velocity.

At the same time that we were making our first observations Keith B. Jefferts, Arno A. Penzias and Robert W. Wilson of the Bell Laboratories were observing DCN, that is, the deuterated form of the hydrogen cyanide molecule (HCN). With the 36-foot radio telescope of the NRAO located on Kitt Peak in Arizona, Jefferts, Penzias and Wilson turned their attention to the Kleinmann-Low nebula, a small cloud in the Great Nebula in Orion in which many interstellar molecules have been detected. Within the Kleinmann-Low nebula the DCN molecule was radiating very strong spectral lines at the wavelengths of two millimeters and four millimeters. The observations indicated that the ratio of DCN to HCN was 1 : 170.

This ratio of DCN to HCN is much higher than the ratio of HDO to H_2O in the earth's oceans. It was immediately realized, however, that if the result was to be understood, certain facts of chemistry would have to be taken into ac-



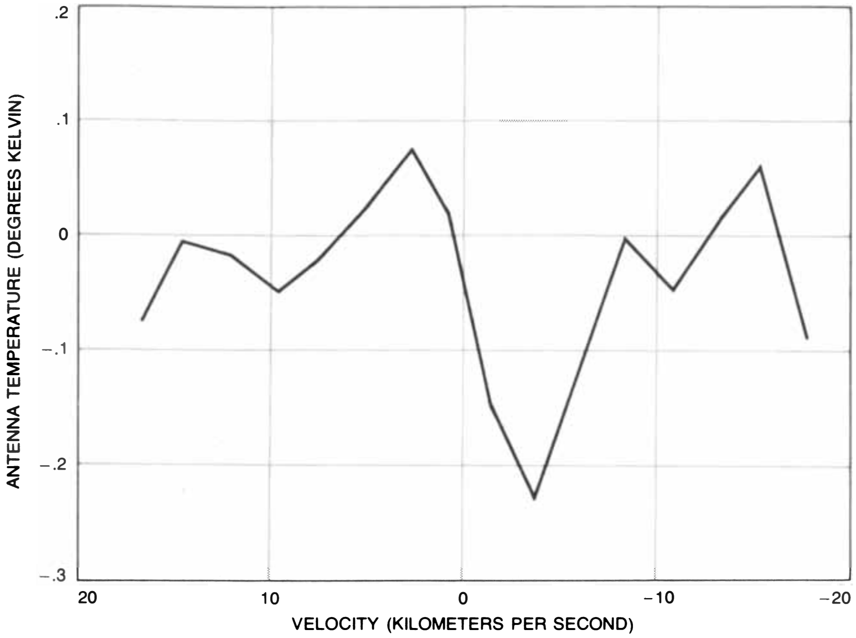


ABUNDANCE OF DEUTERIUM IN THE UNIVERSE as measured both in interstellar space (*black*) and in the solar system (*color*) is shown with respect to the abundance of ordinary hydrogen. For each measurement of the abundance of deuterium made in the solar system and in interstellar space the names of the investigators, the region or the direction observed and the form of deuterium detected are listed in that order. Detailed information about most of the investigators and their work is given in the text. The exceptions are the following: David C. Black of the Ames Research Center of the National Aeronautics and Space Administration deduced the abundance of deuterium in the protosolar nebula from which the sun evolved from measurements of deuterated water (HDO) in meteorites (water in which one deuterium atom replaces one of the two hydrogen atoms in H_2O). Mark A. Allen and Richard Crutcher are both of the California Institute of Technology. Dennis J. Hegyi of the Bartol Research Foundation collaborated with Nathaniel P. Carleton and Wesley A. Traub in their measurements of deuterium in the Great Nebula in Orion at visible wavelengths. Harmon Craig is of the University of California at San Diego. G. Boato was at the University of Chicago. Nicola Grevesse is at the Institute of Astrophysics at Liège in Belgium. "Recombination line" refers to a spectral line emitted at a radio wavelength when an electron recombining with a deuterium ion to form un-ionized deuterium passes from the 93rd to the 92nd energy level. "Lyman lines" are the spectral lines emitted by a hydrogen or deuterium atom when electron drops from higher energy states to ground state. "From ^3He" refers to an abundance of deuterium deduced from the abundance of helium 3.

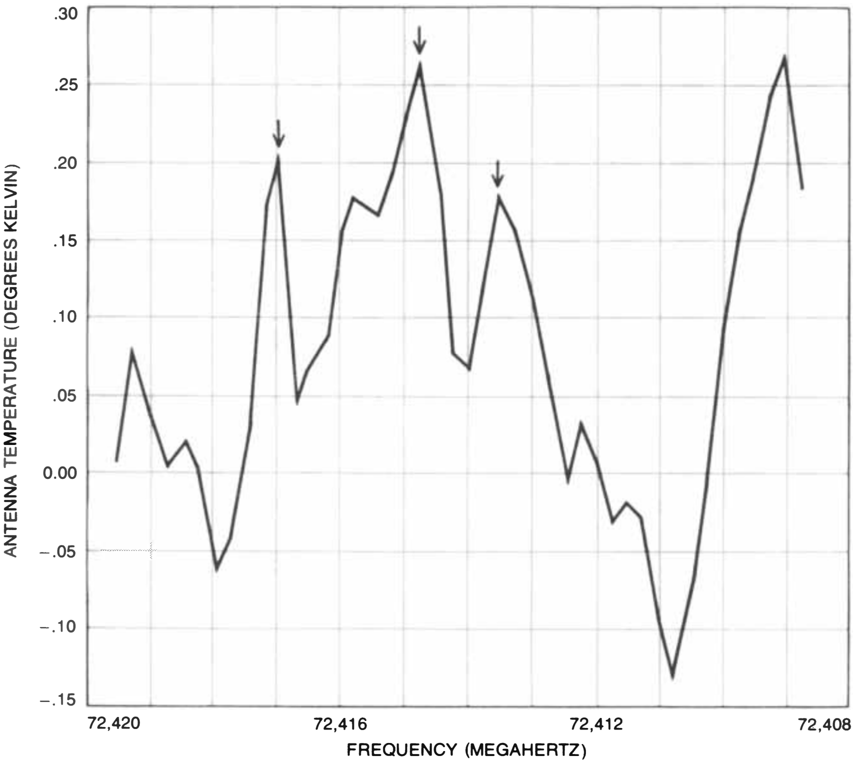
count in addition to the facts of physics. Chemical combinations of the elements follow complicated principles and processes, many of which are unknown. That is particularly true for elements combining on the surface of dust grains in interstellar clouds, which is a leading possibility for the way in which such molecules form. The ratio of deuterium to ordinary hydrogen in interstellar space could very well be quite different from the ratio of DCN to HCN. Calculations show that the difference in the ratios from such chemical fractionation may be on the order of a factor of 600. When the observation of Jefferts, Penzias and Wilson was adjusted to take the correction factor into account, the ratio of deuterium to ordinary hydrogen in the Kleinmann-Low nebula was found to be 1 : 100,000, just below the range deduced from the observations of Cesarsky, Moffet and Pasachoff. The fact that the ratio of deuterium to hydrogen is indeed substantially lower than the ratio of DCN to HCN has since been confirmed by two independent limits set on the abundance of deuterium by Cesarsky observing at radio wavelengths and by another group observing a transition of deuterium at visible wavelengths.

The abundance ratio of deuterium to ordinary hydrogen has also been determined from space, from the third Orbiting Astronomical Observatory satellite, named *Copernicus*. One experiment, conducted by a group from Princeton University including Lyman Spitzer, Jr., Jerry F. Drake, Edward B. Jenkins, Donald C. Morton, John B. Rogerson, Jr., and Donald G. York, uses a 32-inch telescope to observe spectra at the ultraviolet wavelengths between 950 and 1,450 angstroms and between 1,650 and 3,000 angstroms. The second range includes spectral lines from molecular hydrogen (H_2), detected only once before, in an experiment placed aboard a sounding rocket. Molecular hydrogen had not been detected from the earth's surface because it has no lines in the visible region of the spectrum, and its radiation in the ultraviolet region is absorbed by the atmosphere.

The results from the Princeton experiment aboard *Copernicus* showed that whenever the telescope was pointed toward reddened stars (that is, in directions where the interstellar material in front of stars affects the overall distribution of the energy radiated by the stars by favoring the longer—redder—wavelengths), the spectra revealed that at least 10 percent of the intervening mat-



ABSORPTION LINE OF DEUTERIUM shows up as the central dip in observations of the galactic center at a wavelength of 92 centimeters. The observations were made with the 130-foot radio telescope at the Owens Valley Radio Observatory in California. The curve is the sum of results from observations made in 1972 and 1973. The dip is shifted in frequency by a small amount, indicating that the gas incorporating the deuterium is moving at a velocity of 3.7 kilometers per second toward the solar system. Spectrum is calibrated in terms of temperature of antenna, equivalent to energy of radiation received by antenna.



EMISSION LINE OF DEUTERATED HYDROGEN CYANIDE (DCN), which is actually composed of three separate lines (*arrows*), was observed by Keith B. Jefferts, Arno A. Penzias and Robert W. Wilson of the Bell Laboratories at millimeter wavelengths with the 36-foot radio telescope of the National Radio Astronomy Observatory on Kitt Peak in Arizona. The molecule was detected in Kleinmann-Low nebula in constellation Orion.

ter was in the form of molecular hydrogen. The same result was obtained with 11 different stars. For a similar number of unreddened stars, however, molecular hydrogen was not detected at all to a limit of one part in 10 million.

Furthermore, the Princeton group measured two lines of HD (deuterated H_2) at wavelengths of 1,054 and 1,066 angstroms. The first results, for nine stars, indicate that the abundance ratio of HD to H_2 is 1 : 1,000,000. Again the result had to be adjusted by calculations of chemical combinations to yield a corrected ratio of deuterium to ordinary hydrogen. John H. Black and Alexander Dalgarno of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory have calculated that the ratio of deuterium to ordinary hydrogen is between 1 : 5,000 and 1 : 500,000 for the nebula Zeta Ophiuchus.

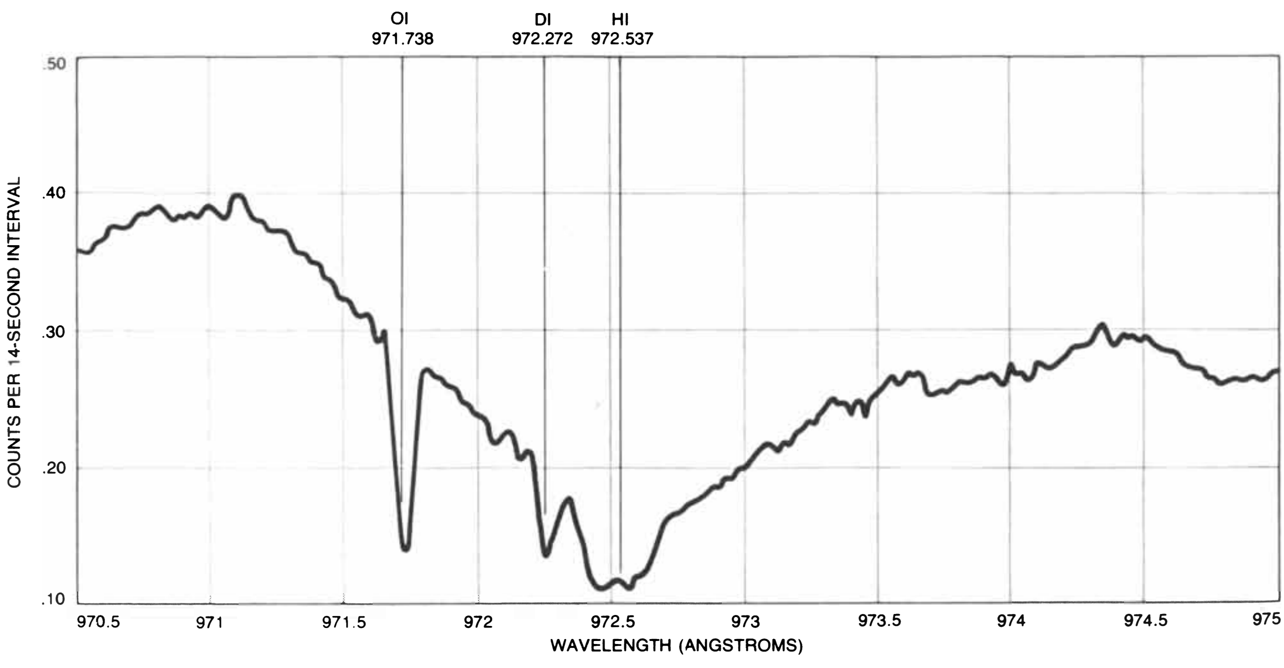
After these promising measurements had been made Rogerson and York attempted to use the telescope aboard *Copernicus* to observe the Lyman series of transitions of the deuterium atom. These transitions absorb radiation in the ultraviolet region of the spectrum when the electron in the deuterium atom is raised to a higher energy level from the "ground" state. The wavelengths of the transitions are slightly shorter than the wavelengths of the corresponding transitions for the atom of ordinary hydrogen (and were also slightly shorter than the wavelengths for which the experiment was actually designed). The calculation of the ratio of deuterium to ordinary hydrogen from these measurements is straightforward and not subject to the correction factors that must be applied to the molecular observations.

Rogerson and York searched for the transitions in the interstellar gas in front of hot stars of spectral Type B. Such stars have few lines in their own spectra and therefore any absorption lines detected would have been formed in interstellar space between the star and the earth. In the direction of the star Beta Centauri, Rogerson and York were able to measure four lines in the Lyman series for deuterium. The ratio of deuterium to ordinary hydrogen in that direction is 1 : 70,000, with an error of 15 percent. The work is being continued for other stars.

The big-bang theory assumes that the universe is isotropic (that it looks more or less the same in all directions), so that a knowledge of how uniformly deuterium is distributed through space bears on whether or not the big bang is indeed a good model for the origin of the universe. If the abundance of deuterium is found to be nonuniform, for example by comparing the observations from *Copernicus* of nearby stars with our observations toward the galactic center, then either the big bang is not a good model or the deuterium was formed in some way other than the big bang.

In order for the measurements of deuterium to be significant in cosmological calculations, the measurements must reflect a general cosmic abundance and not simply local variations. For this reason it is difficult to interpret the abundance of deuterium in the solar system in terms of a big-bang origin. That abundance should nonetheless be discussed briefly if only because deuterium had not been detected elsewhere in the solar system until last year.

Some deuterium has been found in the water content of carbonaceous meteorites. Until the Apollo astronauts landed on the moon, meteorites were the only samples of matter from space. One of the Apollo experiments was designed to capture ionized atoms from the solar wind: the flux of ions expelled from the sun. The astronauts caught the ions in a "window shade" of aluminum foil they unrolled on the moon. The foil was then brought back to earth for analysis. Johannes Geiss of the University of Bern and Hubert Reeves of the Nuclear Research Center at Saclay and the Institute of Astrophysics in Paris made deductions about the abundance of deuterium from the light isotope helium 3 that had



SPECTRUM OF THE STAR BETA CENTAURI in the ultraviolet region shows Lyman-gamma absorption lines for ordinary hydrogen (HI) and deuterium (DI). The lines are from the interstellar gas between the star and the solar system and they are superposed on a broad Lyman-gamma line from the star (*gentle dip*). They were detected with the ultraviolet spectrometer aboard the third Orbiting Astronomical Observatory, named *Copernicus*. A line of oxygen (OI) also happens to be in this region of the spectrum.

been found in the foil. It is known that the sun has transformed into helium 3 most of the deuterium that was present in the primordial solar nebula, the cloud of dust and gas out of which the sun formed. Therefore the amount of helium 3 in the foil directly reflects an upper limit for the amount of deuterium that could have been present in the solar nebula. Geiss and Reeves have calculated that the abundance ratio of deuterium to ordinary hydrogen in the solar nebula was 1 : 40,000.

It is also possible that the deuterium in the material from which the planets were made was not transformed into helium 3 as the protosun materialized out of the solar nebula. In March, 1972, Reinhard Beer of the Jet Propulsion Laboratory of the California Institute of Technology and his collaborators reported that they had detected CH_3D, a deuterated form of methane (CH_4), in the spectrum of Jupiter at infrared wavelengths. Beer and Frederic W. Taylor calculated the molecular-correction factors and found that the abundance ratio of deuterium to ordinary hydrogen in Jupiter was between 1 : 13,000 and 1 : 35,000, depending on certain assumptions about the structure of the planet's atmosphere.

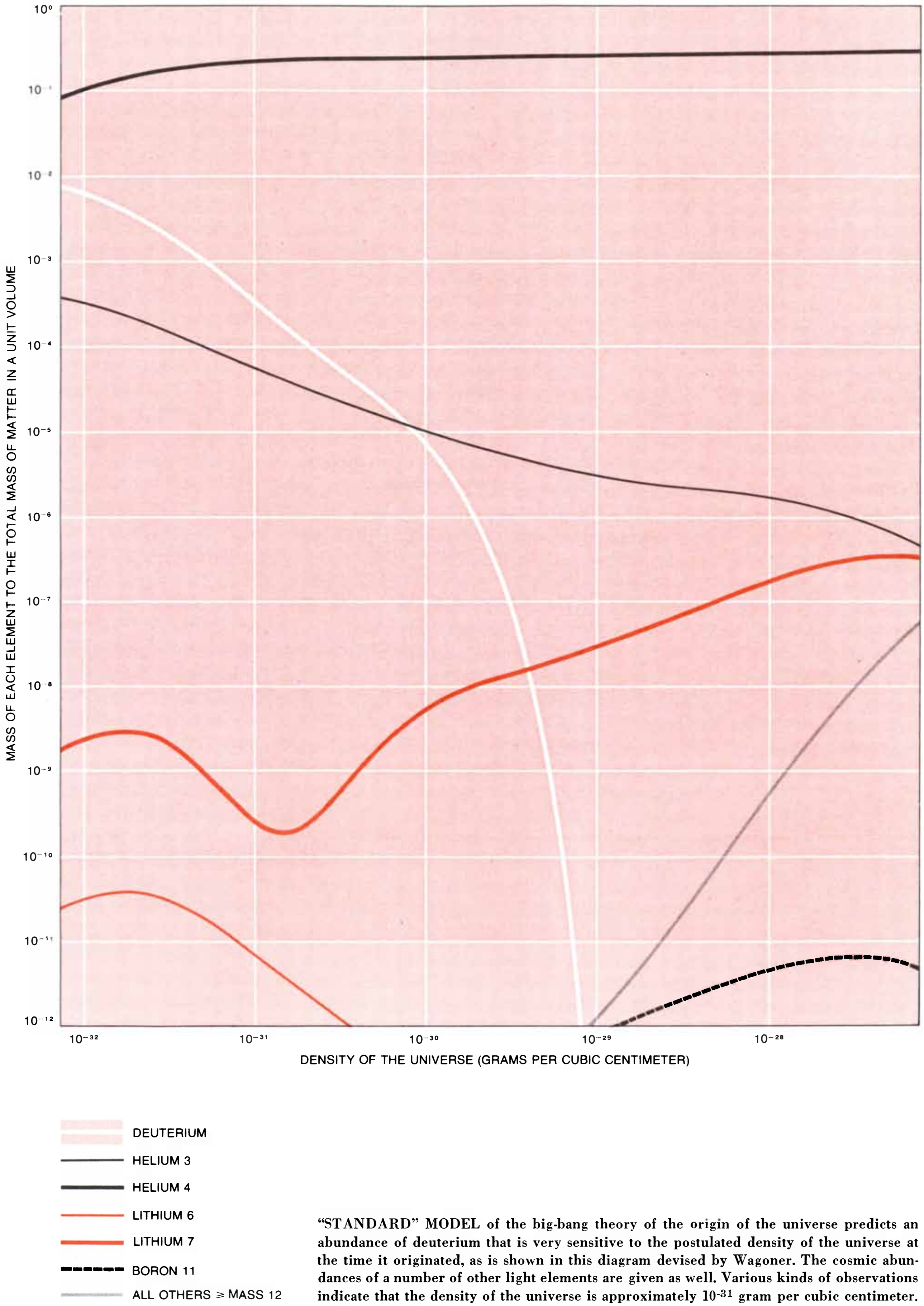
Although the interpretation of results from such a relatively complicated molecule as CH_3D is very uncertain, the calculations seem to show a clear discrepancy with the value of 1 : 6,600 found on the earth. Within the past year John T. Trauger and Frederick L. Roesler of the University of Wisconsin and Nathaniel P. Carleton and Wesley A. Traub of the Center for Astrophysics have detected three lines of deuterated molecular hydrogen (HD) on Jupiter in the near-infrared region of the spectrum. From their observations they calculate a ratio of deuterium to ordinary hydrogen of 1 : 48,000, with a possible error of 20 percent. The discrepancy between the values on the earth and those on Jupiter might be due to the fact that the ratio of heavy water (HDO) to ordinary water (H_2O) on the earth could be enhanced over the ratio of deuterium to ordinary hydrogen by some chemical fractionation process.

A most unusual observation of deuterium was made two years ago. A group of workers from the University of New Hampshire were observing the sun during the large solar flares of August, 1972, with a gamma ray instrument aboard Orbiting Solar Observatory 7. During the particularly large flare of August 7 they detected a peak in the gamma ray spectrum corresponding to the energy that would have been released by the formation of deuterons. That peak meant deuterium was being synthesized on the surface of the sun. This event, however, is probably unrelated to the total cosmic abundance of deuterium. The amount of deuterium formed in this way is small, and it may well be quickly consumed by further nuclear processes on the sun before it can be expelled into space.

The results of the observations of deuterium to date are summarized in the table on page 112. What do these indications of the amount of deuterium in interstellar space tell us about the early universe? One can link the ratio of deuterium to ordinary hydrogen to the big bang by theoretical calculations. For this purpose we shall adopt the standard model of the big bang developed by Wagoner, Fowler and Hoyle, which accepts three basic assumptions about the universe today: first, the universe is homogeneous and isotropic; second, the principle of equivalence must hold, that is, a gravitational field cannot be distinguished from an acceleration (a principle fundamental to the theory of relativity); third, the three-degree background radiation was indeed generated by the big bang. The model also makes several other assumptions. The baryon number must be positive, that is, the amount of antimatter is not equal to the amount of matter, ruling out theories to the contrary. The lepton number is small, that is, the flux of neutrinos does not overwhelm the amount of radiation. The general theory of relativity provides the correct theory of gravitation. Lastly, only the kinds of subatomic particles that we now know about are present; there are no new ones.

In the standard model it happens that the amount of deuterium formed just after the big bang is sensitive to the den-

KLEINMANN-LOW NEBULA is embedded in Great Nebula in Orion. Location, which is indicated by contours, is rich in many molecules, including deuterated hydrogen cyanide.



"STANDARD" MODEL of the big-bang theory of the origin of the universe predicts an abundance of deuterium that is very sensitive to the postulated density of the universe at the time it originated, as is shown in this diagram devised by Wagoner. The cosmic abundances of a number of other light elements are given as well. Various kinds of observations indicate that the density of the universe is approximately 10^{-31} gram per cubic centimeter.

sity of the universe at that time. If the density were relatively high in the first few seconds after the big bang, the deuterium just formed would have been quickly cooked into helium and the end result would be less deuterium. Conversely, if the universe was less dense, the end result would have been more deuterium [*see illustration on the opposite page*]. Thus from the measured value of the deuterium in the universe today it is possible to find the density of the universe at the time that the elements were first synthesized. Knowing the density of the universe then and the rate of expansion now, it is possible to calculate what the density of the universe is at present. The density we deduce from our observations is of the order of 10^{-31} gram per cubic centimeter. That density is not enough to "close" the universe, which will thus continue to expand indefinitely. Therefore knowing the universe's present density enables us to draw conclusions about its present state and eventual future.

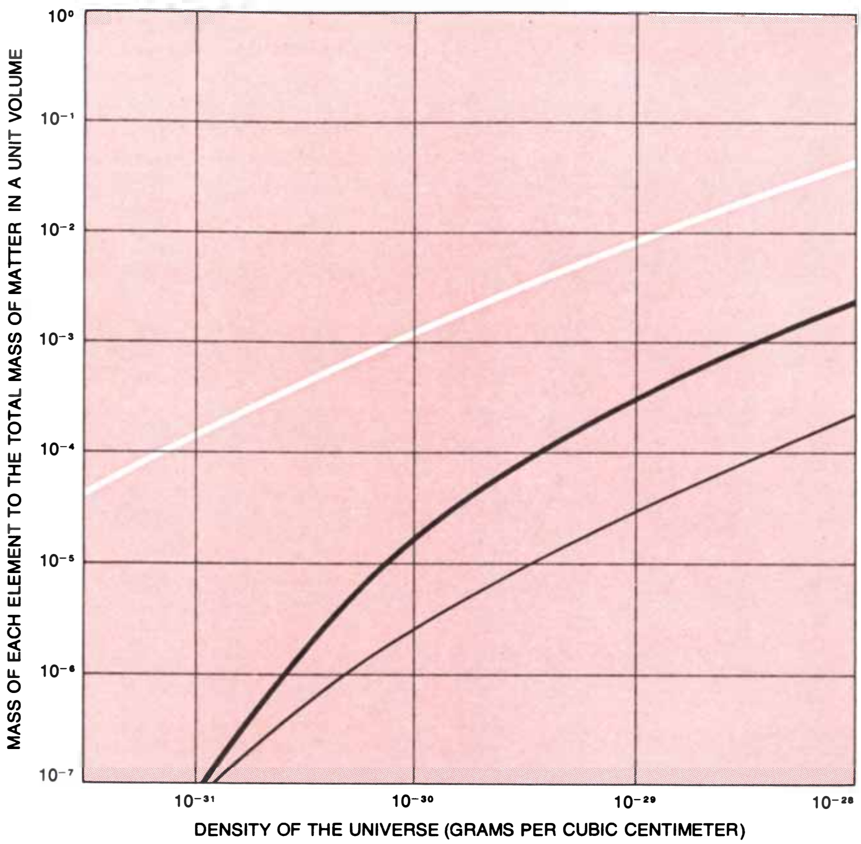
In addition to the standard model of the big bang there is a very speculative new hypothesis called the statistical-bootstrap model, advanced by Robert D. Carlitz of the University of Chicago, Steven C. Frautschi of Cal Tech and Werner Nahm of the University of Bonn. It assumes that massive "superbaryons" formed just after the big bang. The superbaryons were approximately the same size as ordinary baryons, but their mass was as much as 10^{38} times greater. At that point the force of gravity becomes competitive with the normal forces that govern the interaction of nuclear particles.

The statistical-bootstrap model assumes that there are many elementary particles that have not yet been observed because the energies required to make them cannot be attained in the laboratory. The model also assumes that the laws governing the existence of such particles, which have been determined in the laboratory for particles with up to 400 times the mass of an ordinary baryon, hold true for a further factor in mass of 10^{35}! In one sense this new hypothesis, in which there are many unknown particles, is the opposite of the standard theory, which assumes that there are none. There is, in fact, no a priori reason to favor one theory over the other.

If the statistical-bootstrap model turns out to be accurate, there will be many consequences for cosmology. First, no helium would have been formed in the early universe because no nuclear particles would have been available for a full year after the big bang, and by that time the density of the universe would have been too low for helium to be synthesized. Wagoner has calculated the effect of the statistical-bootstrap model on the formation of deuterium. Contrary to the predictions of the standard model, the statistical-bootstrap model predicts that a less dense universe would result in a lower abundance of deuterium [*see illustration above*].

There are two other ways to measure the density of the universe. The first entails simply adding up the masses of everything that can be observed (stars, galaxies and so forth) and dividing the total mass in a given unit volume by that volume. The most recent work along this line has been done by Stuart L. Shapiro of Princeton. His results depend somewhat on the value accepted for the Hubble constant, that is, on the rate at which the universe is expanding. The value of the Hubble constant is currently a topic of debate in some circles. The traditional value is 75 kilometers per second per megaparsec, which corresponds to a density of 10^{-31} gram per cubic centimeter. (One megaparsec is 3.3 million light-years.) A newer value measured by Allan R. Sandage of the Hale Observatories and Gustav Tammann of the University of Basel is 55 kilometers per second per megaparsec, which yields three-fourths the density of the former value.

The second method of measuring the density of the universe considers the dynamics of interactions of clusters of galaxies; it predicts a density that is substantially higher than the value obtained by the first method. Both values for the density could be correct if there is a substantial amount of invisible material in the universe. This "missing mass" could be in the form of molecular hydrogen in the intergalactic material, although the measurements from the *Copernicus* sa-



"STATISTICAL BOOTSTRAP" MODEL of the big-bang theory, advanced by Robert D. Carlitz of the University of Chicago, Steven C. Frautschi of the California Institute of Technology and Werner Nahm of the University of Bonn, predicts a relation between the abundance of deuterium (*white line*) and the density of the universe quite different from what is predicted by the standard model. At the density of 10^{-31} gram per cubic centimeter the abundance of helium 3 (*thin black line*) and of ordinary helium (*thick black line*) is considerably less than is predicted by the standard model or observed. The density of 10^{-31} gram per cubic centimeter is insufficient to "close" the universe, that is, to stop its expansion.

tellite seem to indicate that there is not enough molecular hydrogen to account for it. Alternatively, the mass could be in the form of the enigmatic black holes. The observed value for the abundance of deuterium in the universe can place an independent limit on how much invisible matter there could be.

The amount of missing mass in the universe touches on a major question in cosmology: Will the universe go on expanding forever, or is there enough matter in it for mutual gravitational attraction to eventually reverse the expansion? In cosmological terms, is the amount of missing mass large enough to "close" the universe? There is a parameter q_0 that is used in cosmological equations. It is inversely proportional to the Hubble constant, and it represents the rate at which the expansion of the universe is slowing down. If q_0 equals 0, the expansion is not slowing down and the universe is open. If q_0 equals 1, then the universe is closed and will eventually collapse on itself. The dividing line between an open universe and a closed universe is a q_0 of 1/2. In the steady-state model of the universe q_0 is -1.

Sandage has published a "formal" value for q_0 of $.96 \pm .4$, which would mean that the expansion is slowing down and the universe is closed. His value is based on direct observation of the distant galaxies. The uncertainty in the value, however, means that a q_0 of 1/2 is not ruled out; $.96 - .4$ is .56, or virtually 1/2. On that basis the universe would be open. If the deuterium originated in the big bang with an abundance ratio of the levels now being detected, then q_0 is less than .1 and the universe would definitely be open. It is interesting that according to Wagoner's calculations that is the case for both the standard model and the statistical-bootstrap model. There are other cosmological methods of assessing q_0 that also yield low values for the density.

There are several ways to get around the possible disagreement between Sandage's value of q_0 from observations of the distant galaxies and Wagoner's value of q_0 from measurements of the abundance of deuterium. Beatrice Tinsley of the University of Texas has hypothesized that because galaxies evolve over a period of time, an observed value of q_0 equal to 1 could mean that the real value of q_0 is 0. The essential point of her argument is that as we look deeper into space we are also looking backward in time, and that the average brightness of galaxies long ago may be quite different from what it is today. Such a fact would distort the scale of distances, which is based on galaxies having a unique average brightness, and thus the value of q_0 would also be distorted. Tinsley's proposal is a controversial one. Sandage has shown that in order for it to account for the discrepancy between his observations and Wagoner's calculations, the galaxies would have to be decreasing in brightness by about .1 magnitude per billion years, which is very fast according to many theories of the evolution of galaxies.

Another way to resolve the disagreement would be to accept the notion that significant amounts of deuterium were formed after the big bang. Recently one of us (Fowler) and Hoyle have calculated that deuterium could have formed in space by a spallation reaction: if a shock wave could inject helium nuclei into an ionized gas composed mainly of hydrogen, deuterium could be knocked out of the helium nuclei by the force of their impact on the nuclei of ordinary hydrogen. Alternatively, if only a neutron were knocked out of the helium nucleus, the neutron could combine with a proton in the surrounding gas to form a deuteron. Deuterium that came about in this way would not readily break up again; since the temperature of the gas would be low, the amount of energy in it would not be large enough to cause the nuclei to dissociate. Exactly how important such a spallation process would be on the galactic scale remains to be seen. Possibly the process might enhance the ratio of deuterium to ordinary hydrogen in certain local regions such as in the Orion nebula and not in others such as the galactic center.

Another mechanism for producing deuterium apart from the big bang has been proposed by Stirling A. Colgate of the New Mexico Institute of Mining and Technology. The observed deuterium could have been synthesized if 3 percent of the mass of the galaxy had at some time been recycled in supernovas. Deuterium could also have been synthesized in supermassive stars, as was suggested by Fowler and Hoyle. Accepting the hypothesis that deuterium was formed apart from the big bang would probably require a number of significant changes in our views of the relative importance of such phenomena as supernovas in galaxies.

Cosmological problems never seem to admit of easy solutions. With the current observational and theoretical work on deuterium moving ahead so rapidly, however, new lines of reasoning and inquiry are opening up for tackling the most basic questions of the universe.

MATHEMATICAL GAMES

On the contradictions of time travel, and answers to last month's problems

by Martin Gardner

"It's against reason," said Filby.

"What reason?" said the Time Traveller.

—H. G. WELLS, *The Time Machine*

H. G. Wells's short novel *The Time Machine,* an undisputed masterpiece of science fiction, was not the first story about a time machine. That distinction belongs to "The Clock That Went Backward," a pioneering but mediocre yarn by Edward Page Mitchell, an editor of the New York *Sun.* It was published anonymously in the *Sun* on September 18, 1881, seven years before young Wells (he was only 22) wrote the first version of his famous story.

Mitchell's tale was so quickly forgotten that science-fiction buffs did not even know of its existence until Sam Moskowitz reprinted it last year in his anthology of Mitchell's stories, *The Crystal Man.* Nor did anyone pay much attention to Wells's fantasy when it was serialized in 1888 in *The Science Schools Journal* under the horrendous title "The Chronic Astronauts." Wells himself was so ashamed of this clumsily written tale that he broke it off after three installments and later destroyed all the copies he could find. A completely rewritten version, "The Time Traveller's Story," was serialized in *The New Review* beginning in 1894. When it came out as a book in 1895, it brought Wells instant recognition.

One of the many remarkable aspects of Wells's novella is the introduction in which the Time Traveller (his name is not revealed, but in Wells's first version he is called Dr. Nebo-gipfel) explains the theory behind his invention. Time is a fourth dimension. An instantaneous cube cannot exist. The cube we see is at each instant a cross section of a "fixed and unalterable" four-dimensional cube having length, breadth, thickness and duration. "There is no difference between Time and any of the three dimensions of Space," says the Time Traveller, "except that our consciousness moves along it." If we could view a person from outside our space-time (the way human history is viewed by the Eternals in Isaac Asimov's *The End of Eternity* or by the Tralfamadorians in Kurt Vonnegut's *Slaughterhouse-Five*), we would see that person's past, present and future all at once, just as in 3-space we see all parts of a wavy line that traces on a time chart the one-dimensional spatial movements of mercury in a barometer.

Reading these remarks today, one might suppose that Wells had been familiar with Hermann Minkowski's great work of tidying up Einstein's general theory of relativity. The line along which our consciousness crawls is, of course, our "world line": the line that traces our movements in 3-space on a four-dimensional Minkowski space-time graph. (*My World Line* is the title of George Gamow's autobiography.) But Wells's story appeared in its final form 10 years before Einstein published his first paper on relativity!

When Wells wrote his story, he regarded the Time Traveller's theories as little more than metaphysical hanky-panky designed to make his fantasy more plausible. A few decades later physicists were taking such hanky-panky with the utmost seriousness. The notion of an absolute cosmic time, with absolute simultaneity between distant events, was swept out of physics by Einstein's tensor equations. Virtually all physicists now agree that if an astronaut were to travel to a distant star and back, moving at a velocity close to that of light, he could in theory travel thousands of years into the earth's future. Kurt Gödel constructed a cosmological model in which one can in principle travel to any point in the world's past as well as future. In 1965 Richard P. Feynman received a Nobel prize for his space-time approach to quantum mechanics in which antiparticles are viewed as particles momentarily moving into the past.

Hundreds of science-fiction stories have been written about time travel, many of them raising questions about time and causality that are as profound as they are sometimes funny. To give the most hackneyed example, suppose you traveled back to last month and shot yourself through the head. Not only do you know before making the trip that nothing like this happened but, assuming that somehow you could murder your earlier self, how could you exist next month to make the trip? The late Fredric Brown's "First Time Machine" opens with Dr. Grainger exhibiting his machine to three friends. One of them uses the device to go back 60 years and kill his hated grandfather when the man was a youth. The story ends 60 years later with Dr. Grainger showing his time machine to two friends.

It must not be thought that logical contradictions arise only when people travel in time. The transportation of anything can lead to paradox. There is a hint of this in Wells's story. When the Time Traveller sends a small model of his machine into the past or the future (he does not know which), his guests raise two objections. If the time machine went into the future, why do they not see it now, moving along its world line? If it went into the past, why did they not see it there before the Time Traveller brought it into the room?

One of the guests suggests that perhaps the model moves so fast in time it becomes invisible, like the spokes of a rotating wheel. But what if a time-traveling object stops moving? If you have no memory of a cube on the table Monday, how could you send it back to Monday's table on Tuesday? And if on Tuesday you go into the future, put the cube on the table Wednesday, then return to Tuesday, what happens on Wednesday if on Tuesday you destroy the cube?

Objects carried back and forth in time are sources of endless confusion in certain science-fiction tales. Sam Mines once summarized the plot of his own story "Find the Sculptor" as follows: "A scientist builds a time machine, goes 500 years into the future. He finds a statue of himself commemorating the first time traveler. He brings it back to his own time and it is subsequently set up in his honor. You see the catch here? It had to be set up in his own time so that it would be there waiting for him when he went into the future to find it. He had to go into the future to bring it back so it could be set up in his own time. Somewhere a piece of the cycle is missing. When was the statue made?"

A splendid example of how paradox arises, even when nothing more than messages go back in time, is provided by the recent conjecture that tachyons, particles moving faster than light, might actually exist [see "Particles That Go Faster than Light," by Gerald Feinberg; SCIENTIFIC AMERICAN, February, 1970]. Relativity theory leaves no escape from the fact that anything moving faster than light would move backward in time. This is what inspired A. H. Reginald Butler, a Canadian botanist, to write his often quoted limerick:

There was a young lady named Bright
Who traveled much faster than light.
She started one day
In the relative way,
And returned on the previous night.

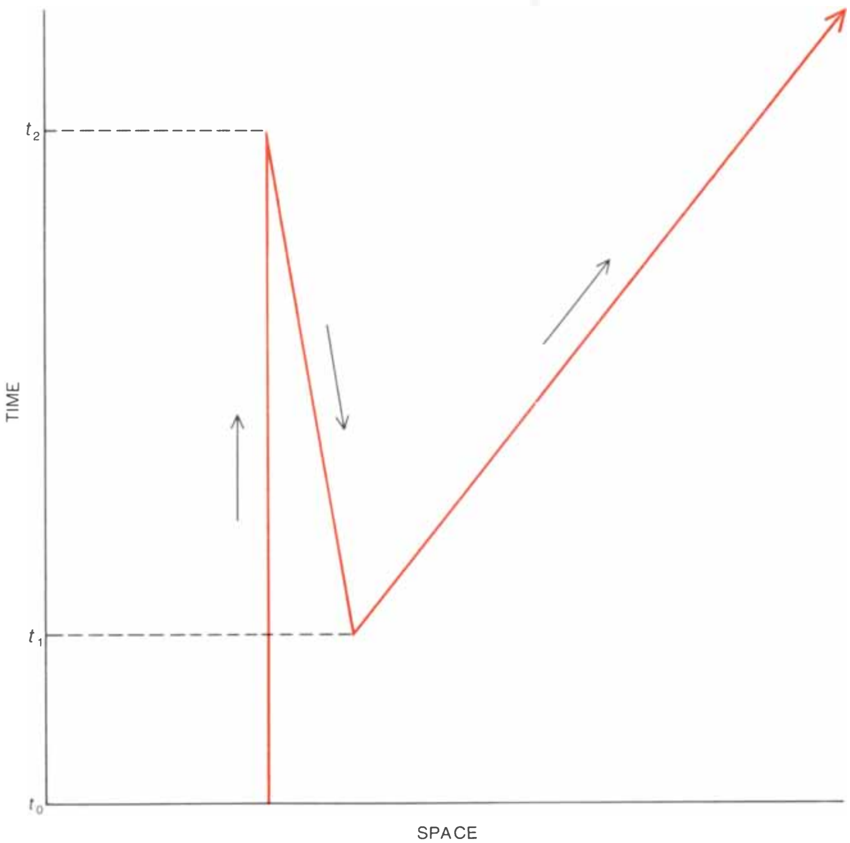
Tachyons, if they exist, clearly cannot be used for communication. G. A. Benford, D. L. Book and W. A. Newcomb (of "Newcomb's paradox," the topic of this department in March) have chided physicists who are searching for tachyons for overlooking this. In "The Tachyonic Antitelephone" (*Physical Review, D*, Volume 2, July 15, 1970, pages 263–265) they point out that certain methods of looking for tachyons are based on interactions that make possible, in theory, communication by tachyons. Suppose physicist Jones on the earth is in communication by tachyonic antitelephones with physicist Alpha in another galaxy. They make the following agreement. When Alpha receives a message from Jones, he will reply immediately. Jones promises to send a message to Alpha at three o'clock earth time, if and only if he has not received a message from Alpha by one o'clock. Do you see the difficulty? Both messages go back in time. If Jones sends his message at three, Alpha's reply could reach him before one. "Then," as the authors put it, "the exchange of messages will take place if and only if it does not take place . . . a genuine . . . causal contradiction." Large sums of money have already gone down the drain, the authors believe, in efforts to detect tachyons by methods that imply tachyonic communication and are therefore doomed to failure.

Time dilation in relativity theory, time travel in Gödel's cosmos and reversed time in Feynman's way of viewing antiparticles are so carefully hedged by other laws that contradictions cannot arise. In most time-travel stories the paradoxes are skirted by leaving out any incident that would generate a paradox. In some stories, however, logical contradictions explicitly arise. When they do, the author may leave them paradoxical to bend the reader's mind or he may try to escape from paradox by making clever assumptions.

Before discussing ways of avoiding the paradoxes, brief mention should be made of what might be called pseudo-time-travel stories in which there is no possibility of contradiction. There can be no paradox, for example, if one simply observes the past but does not interact with it. The electronic machine in Eric Temple Bell's "Before the Dawn," which extracts motion pictures of the past from imprints left by light on ancient rocks, is as free of possible paradox as watching a video tape of an old television show. And paradox cannot arise if a person travels into the future by going into suspended animation, like Rip van Winkle, or Woody Allen in his latest motion picture, *Sleeper,* or the sleepers in such novels as Edward Bellamy's *Looking Backward* or Wells's *When the Sleeper Wakes.* No paradox can arise if one dreams of the past (as in Mark Twain's *A Connecticut Yankee at King Arthur's Court*), or goes forward in a reincarnation, or lives for a while in a galaxy where change is so slow in relation to earth time that when he returns, centuries on the earth have gone by.

But when someone actually travels to the past or the future, interacts with it and returns, enormous difficulties arise. In certain restricted situations paradox can be avoided by invoking Minkowski's "block universe," in which all history is frozen, as it were, by one monstrous space-time graph on which all world lines are eternal and unalterable.

From this deterministic point of view one can allow certain kinds of time travel in either direction, although one must pay a heavy price for it. Hans Reichenbach, in a muddled discussion in *The Philosophy of Space and Time* (Dover, 1957, pages 140–142), puts it this way: Is it possible for a person's world line to "loop" in the sense that it returns him to a spot in space-time, a spot very close to where he once had been and where some kind of interaction, such as speech, occurs between the two meeting selves? Reichenbach argues that this cannot be ruled out on logical grounds; it can only be ruled out on the ground that we would have to give up two axioms that are strongly confirmed by experience: (1) A person is a unique individual who maintains his identity as he ages; (2) a person's world line is linearly ordered so that what he considers "now" is always a unique spot along the line. (Reichenbach does not mention it, but we would also have to abandon any notion of free will.) If we are willing to give up these things, says Reichenbach,

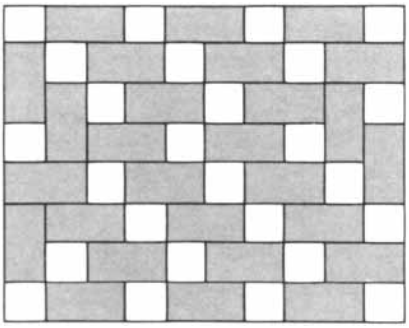


Feynman graph of a time traveler

we can imagine without paradox certain kinds of loops in a person's world line.

Reichenbach's example of a consistent loop is as follows. One day you meet a man who looks exactly like you but who is older. He tells you he is your older self who has traveled back in time. You think him insane and walk on. Years later you discover how to go back in time. You visit your younger self. You are compelled to tell him exactly what your older duplicate had told you when you were younger. Of course, he thinks that you are insane. You separate. Each of you leads a normal life until the day comes when your younger self makes the trip back in time.

Hilary Putnam, in "It Ain't Necessarily So" (*The Journal of Philosophy,* Volume 59, October 25, 1962, pages 658–671), argues in similar fashion that such world-line loops need not be contradictory. He draws a Feynman graph [*see illustration on preceding page*] on which particle pair-production and pair-annihilation are replaced by person pair-production and pair-annihilation. (On Feynman graphs see my article "Can Time Go Backward?" in SCIENTIFIC AMERICAN, January, 1967.) The colored, zigzag line is the world line of time-traveler Smith. At time t_2 he goes back to t_1, converses with his younger self, then continues to lead a normal life. How would this be observed by someone whose world line is normal? Simply put a ruler at the bottom of the chart, its edge parallel to the space axis, and move it slowly upward. At t_0 you see young Smith. At t_1 an older Smith suddenly materializes out of thin air in the same room along with an anti-Smith, who is seated in his time machine and living backward. (If he is smoking, you see his cigarette butt lengthen into a whole cigarette, and so on.) Perhaps the two forward Smiths converse. Finally, at t_2 young Smith, backward Smith and the backward-moving time machine vanish. The older Smith and his older time machine continue on their way. The fact that we can draw a space-time diagram of these events, Putnam insists, is proof that they are logically consistent.



A solution to the Gunport Problem

It is true that they are consistent, but note that Putnam's scenario, like Reichenbach's, involves such weak interaction between the Smiths that it evades the deeper contradictions that arise in time-travel fiction. What happens if the older Smith kills the younger Smith? Will Putnam kindly supply a Feynman graph?

There is only one good way out, and science-fiction scribblers have been using it for at least 40 years. According to Sam Moskowitz, the device was first explicitly employed to resolve time-travel paradoxes by David R. Daniels in "Branches of Time," a tale that appeared in *Wonder Stories* in 1934. The basic idea is as simple as it is fantastic. A person can travel to any point in the future of his universe, with no complications, but the moment he enters the past the universe splits into two parallel worlds, each with its own time track. Along one track rolls the world as if no looping had occurred. Along the other track spins the newly created universe, its history permanently altered. When I say "newly created," I speak, of course, from the standpoint of the time traveler's consciousness. For an observer in, say, a fifth dimension the traveler's world line simply switches from one space-time continuum to another on a graph that depicts all the universes branching like a tree in a metauniverse.

Forking time paths appear in many plays, novels and short stories by non-science-fiction writers. J. B. Priestley uses it in his popular play *Dangerous Corner*. Mark Twain discusses it in *The Mysterious Stranger*. Jorge Luis Borges plays with it in his "Garden of Forking Paths." But it was the science-fiction writers who sharpened and elaborated the concept.

Let us see how it works. Suppose you go back to the time of Napoleon in Universe 1 and assassinate him. The world forks. You are now in Universe 2. If you like, you can return to the present of Universe 2, a universe in which Napoleon had been mysteriously murdered. How much would this world differ from the old one? Would you find a duplicate of yourself there? Maybe. Maybe not. Some stories assume that the slightest alteration of the past would introduce new causal chains that would have a multiplying effect and produce vast historical changes. Other tales assume that history is dominated by such powerful overall forces that even major alterations of the past would damp out and the future would soon be very much the same.

In Ray Bradbury's "A Sound of Thunder" Eckels travels back to an ancient geological epoch under elaborate precautions to prevent any serious alteration of the past. For example, he wears an oxygen mask to prevent his microbes from contaminating animal life. But Eckels violates a prohibition and accidentally steps on a living butterfly. When he returns to the present, he notices subtle changes in the office of the firm that arranged his trip. He is killed for having illegally altered the future. Scores of science-fiction writers have played variations on this theme.

It is easy to see that in this metacosmos of branching worlds it is not possible to produce paradox. The future is no problem. If you travel to next week, you merely vanish for a week and reappear a week younger than you would have been. But if you go back and murder yourself in your crib, the universe obligingly splits. Universe 1 goes on as before, with you vanishing from it when you grow up and make the trip back. Perhaps this happens repeatedly, each cycle creating two new worlds. Perhaps it happens only once. Who knows? In any case Universe 2 with you and the dead baby in it rolls on. You are not annihilated by your deed because now you are an alien from Universe 1 living in Universe 2.

In such a metacosmos it is easy to fabricate hundreds of duplicates of yourself. You can go back a year in Universe 1, live for a year with yourself in Universe 2, then again go back a year to visit two replicas of yourself in Universe 3. Clearly by repeating such loops you can create as many replicas of yourself as you please. They are genuine replicas, not pseudo-replicas as in the scenarios by Reichenbach and Putnam. Each has his independent world line. History might become extremely chaotic, but there is one type of event that can never occur: a logically contradictory one.

This vision of a metacosmos containing branching worlds may seem crazy, but respectable physicists have taken it quite seriously. In Hugh Everett III's Ph.D. thesis " 'Relative State' Formulation of Quantum Mechanics" (*Reviews of Modern Physics,* Volume 29, July, 1957, pages 454–462) he outlines a metatheory in which the universe at every micromicroinstant branches into countless parallel worlds, each a possible combination of microevents that could occur as a result of microlevel uncertainty. The paper is followed by John A. Wheeler's favorable assessment in which he points out that classical physicists were almost as uncomfortable at first with the radical notions of general relativity.

"If there are infinite universes," wrote Fredric Brown in *What Mad Universe,* "then all possible combinations must exist. Then, somewhere, *everything must be true....* There is a universe in which Huckleberry Finn is a real person, doing the exact things Mark Twain described him as doing. There are, in fact, an infinite number of universes in which a Huckleberry Finn is doing every possible variation of what Mark Twain *might* have described him as doing.... And infinite universes in which the states of existence are such that we would have no words or thoughts to describe them or to imagine them."

What if the universe never forks? Suppose there is only one world, this one, in which all world lines are linearly ordered and objects preserve their identity come what may. Brown considers this possibility in his story "Experiment." Professor Johnson holds a brass cube in his hand. It is six minutes to three o'clock. At exactly three, he tells his colleagues, he will place the cube on his time machine's platform and send it five minutes into the past.

"Therefore," he remarks, "the cube should, at five minutes before three, vanish from my hand and appear on the platform, five minutes before I place it there."

"How can you place it there, then?" asked one of his colleagues.

"It will, as my hand approaches, vanish from the platform and appear in my hand to be placed there."

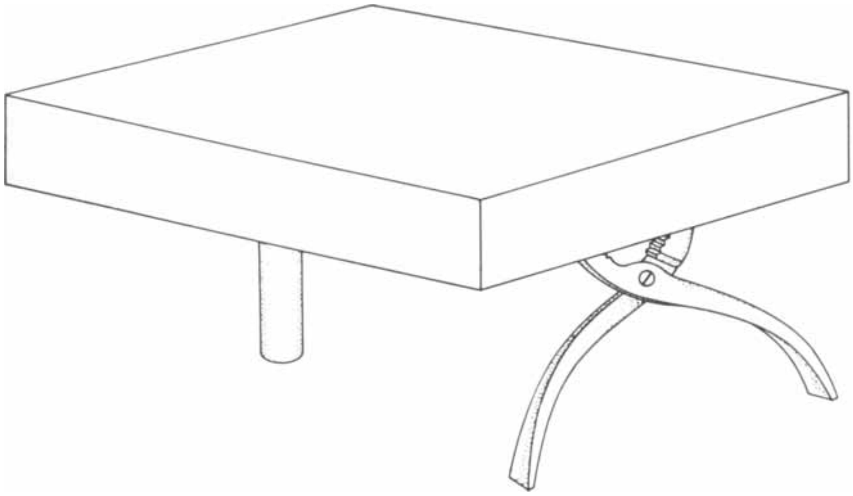
At five minutes to three the cube vanishes from Professor Johnson's hand and appears on the platform, having been sent back five minutes in time by his future action of placing the cube on the platform at three.

"See? Five minutes before I shall place it there, it *is* there!"

"But," says a frowning colleague, "what if, now that it has already appeared five minutes before you place it there, you should change your mind about doing so and *not* place it there at three o'clock? Wouldn't there be a paradox of some sort involved?"

Professor Johnson thinks this is an interesting idea. To see what happens he does not put the cube on the platform at three.

There is no paradox. The cube remains. But the entire universe, including Professor Johnson, his colleagues and the time machine, disappears.



Supporting the board

Last month I presented a set of miscellaneous short problems and promised to give the answers this month. Here they are:

1. The illustration on the opposite page shows one way of placing 27 dominoes on an 8-by-10 field to form 26 holes. Found by Capt. John C. Huval, it was published in *Mathematics Magazine* for November, 1972. Many trivial variations can be produced by sliding one domino, by switching two adjacent dominoes or by rotating a 3-by-3 pattern of three dominoes.

2. The 22 triplets that can be substituted for 7, 13 and 28 in Irvin S. Cobb's story are

$12 \div 2 = 15$,	$15 \div 3 = 14$,	$16 \div 4 = 13$,
$14 \div 2 = 25$,	$18 \div 3 = 24$,	$24 \div 4 = 15$,
$16 \div 2 = 35$,	$24 \div 3 = 17$,	$28 \div 4 = 25$,
$18 \div 2 = 45$,	$27 \div 3 = 27$,	$36 \div 4 = 18$,
$15 \div 5 = 12$,	$18 \div 6 = 12$,	$28 \div 7 = 13$,
$25 \div 5 = 14$,	$36 \div 6 = 15$,	$49 \div 7 = 16$,
$35 \div 5 = 16$,	$48 \div 6 = 17$,	
$45 \div 5 = 18$,		$48 \div 8 = 15$.

Readers interested in how William R. Ransom solved this problem will find it explained in his delightful but little-known book, *One Hundred Mathematical Curiosities* (J. Weston Walch, 1955).

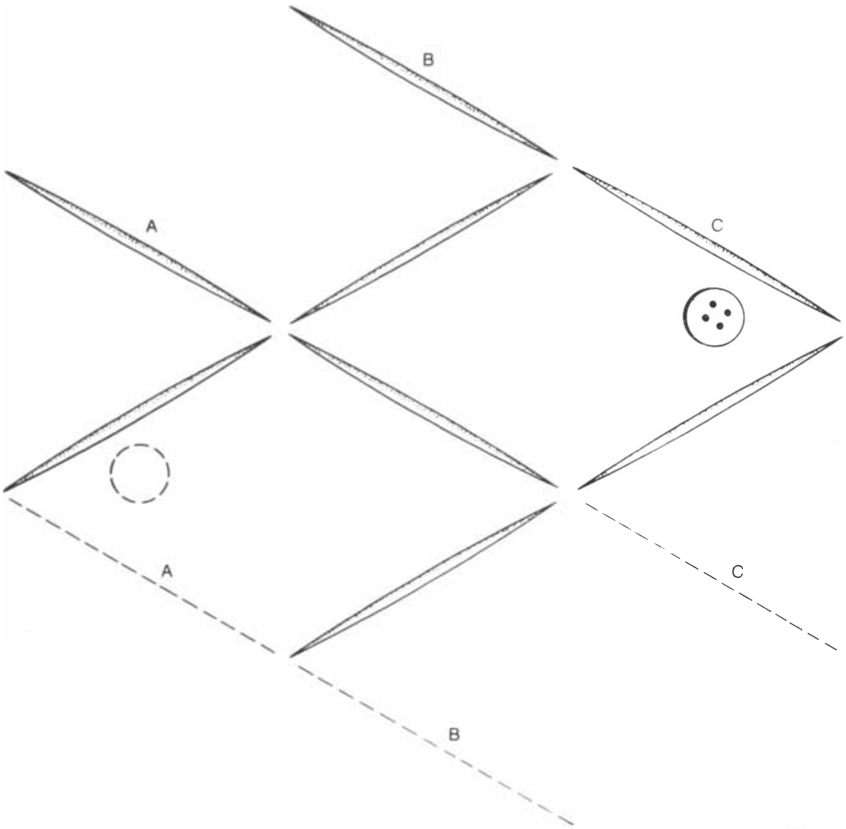
3. The board is supported by the pliers and wooden peg [*see illustration above*]. To tie together the ends of the two hanging cords, tie the magnet to one end and start the cord swinging. Hold the end of the other string and catch the swinging magnet.

4. The key to Raymond Smullyan's monochromatic chess problem lies in the position of the two white pawns. We were told that no piece has moved to a square of a different color, therefore the only way the white king could have escaped from his home square is by castling. The castling must have been on the king's side, otherwise the white rook would have moved from a black square to a white one. If the pawn of unknown color is white, it must have been a rook's pawn that moved to its present square by capturing. But if this was what happened, the white king could not have reached its present position. The rook's pawn, before it made its capture, would have confined the king to KN1 and in its present position would confine the king to KN1 and KR2. Therefore the pawn in question is black.

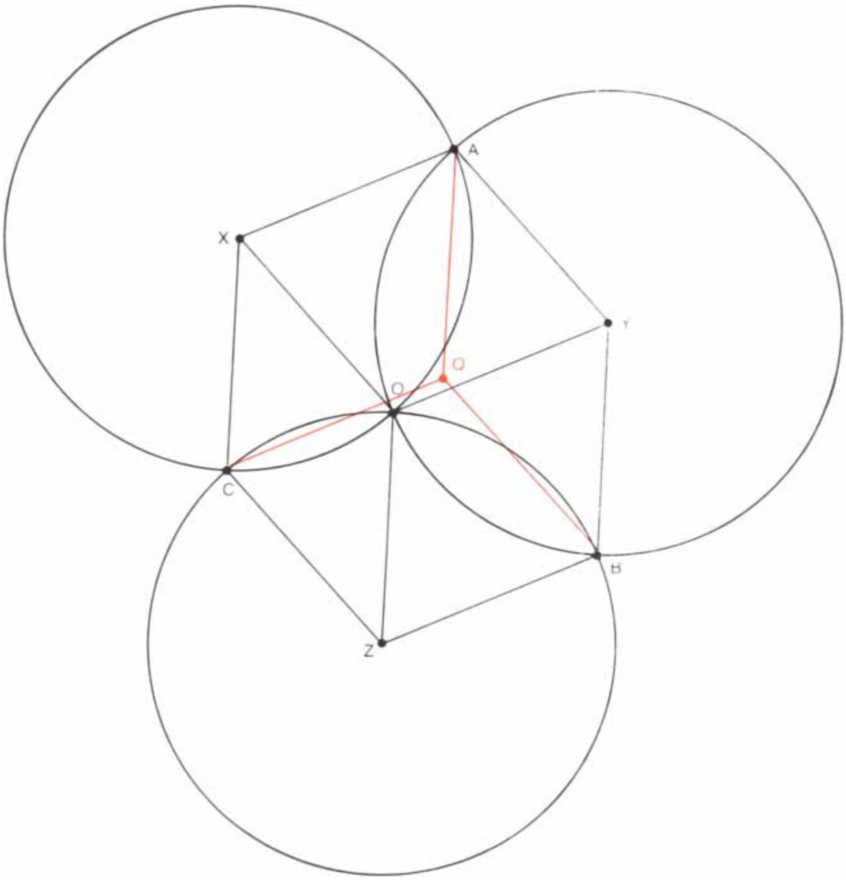
5. Eight swings are enough to reverse those two bookcases. One solution: (1) Swing end *B* clockwise 90 degrees; (2) swing *A* clockwise 30 degrees; (3) swing *B* counterclockwise 60 degrees; (4) swing *A* clockwise 30 degrees; (5) swing *B* clockwise 90 degrees; (6) swing *C* clockwise 60 degrees; (7) swing *D* counterclockwise 300 degrees; (8) swing *C* clockwise 60 degrees.

"If you moved bookcase *AB* in fewer than five swings," writes Robert Abes, who originated this problem, "then you put an end through a wall in mid-swing, or (more likely) wound up with its front side still facing out. If you moved bookcase *CD* without a 300-degree second swing, you either wasted a swing or scooped a hollow out of a wall. Thanks to Jim Lewis for helping me move the large bookcase."

6. We wish to use a penny for decid-



Reversing the fish



Proof of the circle theorem

ing between alternatives A and B with irrational probabilities. The method, devised by Persi Diaconis, applies to any two real fractions.

Notational rules:

a) Express A as an endless binary fraction.

b) Number the digits 1, 2, 3, 4, ... and similarly number the flips of the coin. The nth digit is called the "corresponding digit" of the nth flip.

c) Let the value of each flip be 1 for heads, 0 for tails.

Procedural rules:

a) If the value of a flip equals its corresponding digit, flip again.

b) If the value of a flip is less than its corresponding digit, stop. This decides for A.

c) If the value of a flip is more than its corresponding digit, stop. This decides for B.

Let us see how this works when A is 1/3, B is 2/3. In binary form $A = .01010101\ldots$, $B = .10101010\ldots$. The sequence of flips stops with a decision for A if and only if tails (value 0) appears on a flip whose corresponding digit is 1 in the endless binary fraction for A. The 1's are in even positions, therefore the probability of this happening is $1/2^2 + 1/2^4 + 1/2^6 + \ldots$.

The sum of this series is .01010101.... This is obvious when we consider its binary fractions:

$$\begin{aligned} 1/4 &= .01 \\ 1/16 &= .0001 \\ 1/64 &= .000001 \\ &\vdots \\ &\overline{.010101\ldots} \end{aligned}$$

Similarly, the sequence of flips stops with a decision for B if and only if heads (value 1) appears on a flip corresponding to 0 in the endless fraction for A. The 0's are in odd positions, therefore the probability of this happening is $1/2 + 1/2^3 + 1/2^5 + \ldots$. This is the same as summing $.1 + .001 + .00001 + \ldots$, a series that just as obviously adds to $.10101\ldots = 2/3$.

The specific problem last month was to decide between A equaling the fractional part of π and B equaling 1 minus A. First express A as a binary fraction:

$$A = .00100100001111110110 1\ldots$$

As before, the probability of stopping with a decision for A is the probability that you get a tail (0) on a flip whose corresponding digit is 1. This probability is equal to the binary fraction itself, because the fraction is expressing the probability as the sum of an endless series of binary fractions, each a reciprocal of a power of 2. And the probability of stopping with a decision for B is the probability you will get a head (1) on a flip whose corresponding digit is 0.

In the first case the probability is $1/2^3 + 1/2^6 + 1/2^{11} + \ldots$. (The superscripts are the positions of the 1's in the binary fraction for A.) The sum is .00100100001 . . . , the binary fraction for A.

In the second case the probability is $1/2 + 1/2^2 + 1/2^4 + \ldots$. (The superscripts are the positions of the 0's in the binary fraction for A.) The sum is .1101101111 10 . . . , which is the complement of the previous fraction; that is, 1's have been replaced by 0's and 0's by 1's. It is the binary fraction for B.

It is not hard to see why the method works. Probability A is expressed as a sum of an endless series of probabilities. Each is a disjoint event, so that their sum must equal probability A. There is, of course, a rapidly decreasing probability that the flipping will not stop, but this probability is vanishingly small. The sequence continues only as long as flip values keep matching their corresponding digits. On the nth toss the probability of such a match is 1/2 to the power of n, which has zero measure in the endless series. In other words, the procedure is practically certain to stop, usually quite soon, with a decision.

7. The Jack of Spades.

8. The fish swims the other way if you move three toothpicks and the button as shown in the top illustration on the opposite page.

9. Draw three line segments connecting O to each of the centers, X, Y and Z, of the three circles. Add six line segments to connect each center to the two nearest of the intersection points, A, B and C. The nine line segments are shown in black in the bottom illustration on the opposite page. Each line segment is a unit in length, therefore the lines form three rhombuses. Now through each of the intersection points A, B and C draw another line [*color*], making each line parallel to a radius line segment of one of the circles. This forms three additional rhombuses. Because opposite sides of parallelograms are equal, we know that the three colored line segments are equal and that each is one unit in length. Consequently they meet at a point, Q, which is the center of a circle with a radius of one unit. The intersection points A, B and C lie on this circle, which is the assertion we were asked to prove.

THE AMATEUR SCIENTIST

A plotting device for predicting the orbit of an earth satellite

Conducted by C. L. Stong

Amateurs who are interested in predicting the positions in orbit of the earth satellites that broadcast pictures of cloud cover can now make a device that will perform the calculations automatically. The device is designed for satellites in orbit within 2,000 miles of the earth, and a separate device must be made for each satellite. Details of the apparatus are described by William K. Widger, Jr., who was a pioneer in the development of weather-satellite technology and is now director of geophysical services of Biospherics Consultants International, Inc. (P.O. Box 529, Laconia, N.H. 03246).

"The device consists of a map in the form of a polar projection of the Northern or Southern Hemisphere. A covering sheet of clear plastic is loosely pinned to the center of the map. The pin is fixed at the North or the South Pole.

"The plastic overlay rotates around the pin. The track on the earth's surface over which the satellite passes during any orbit is plotted as a curved line on the overlay [*see illustration on opposite page*]. By rotating the overlay with respect to the polar projection the observer can determine the geographical position of the track, that is, the subpoints on the earth's surface directly under the satellite, at any instant.

"A pattern of concentric circles can be drawn around any point on the earth's surface to represent the local horizon and to serve in determining the series of points directly under a satellite that correspond to the several angles of elevation above the horizon at which the satellite would be seen. When the overlay is rotated to the position at which the track crosses the pattern of concentric circles, the approximate direction and elevation of the satellite are depicted by the points where the track intercepts the circles. The track can be calibrated in terms of the number of minutes that have elapsed since the satellite passed over the Equator on its most recent northbound crossing.

"Normally the observer is interested in the position of the satellite in relation to the observing station. For this reason the map need not be cluttered with outlines of the earth's landmasses, although they can be included if the experimenter wants. They are shown on polar projections of the earth in most world atlases.

"For the background map use polar-coordinate graph paper No. 464417, a product of the Keuffel & Esser Company, or an equivalent. The track of the satellite is drawn with india ink on the type of plastic sheet known as heavyweight clear acetate paper, which is available from dealers in artists' supplies. Mount the graph paper on a stiff backing of cardboard or plywood.

"Begin the project by numbering the major intervals of the graph paper to serve as the geographical coordinates: latitude (ϕ) and longitude (λ). Note that longitude increases both west and east from 0 degrees at the bottom of the map to 180 degrees at the top. Latitude increases from 0 degrees at the Equator to 90 degrees at the Pole. By convention positive numbers designate northern latitudes, negative numbers southern latitudes. In making calculations it is convenient to express longitude in terms of a full circle of 360 degrees, beginning at the bottom of the map and proceeding clockwise. Conversion from one system of designating longitude to the other is easy. Subtract longitudes greater than 180 degrees from 360. Label the difference longitude east. For example, longitude 210 degrees is equivalent to $360 - 210 = 150°$E.

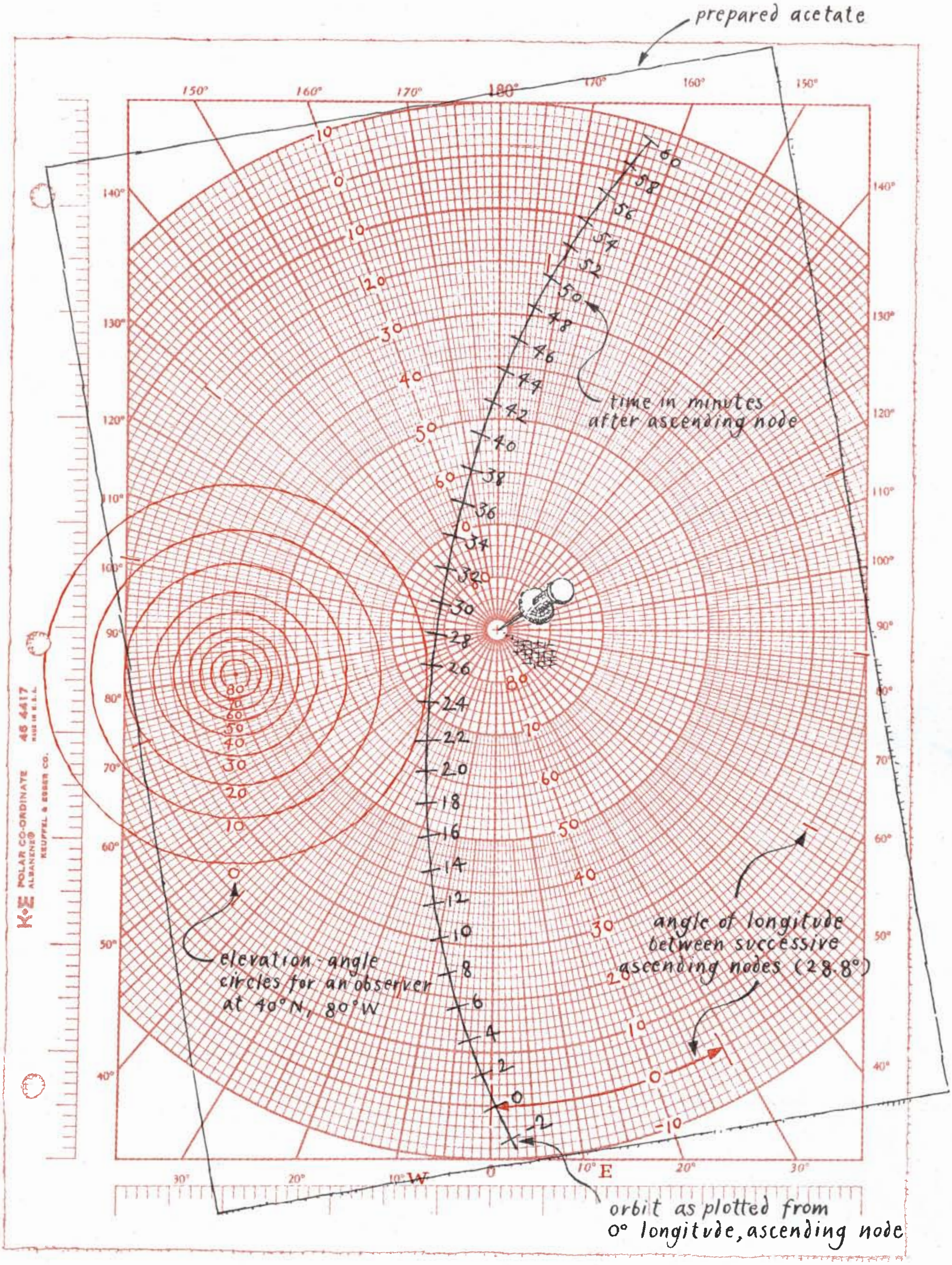
"Next select and identify (by placing a dot on the graph paper) the location where observations will be made. In the accompanying example [*opposite page*] I placed the dot at latitude 40°N and longitude 80°W. I then determined the radii of the concentric circles to be inscribed around this point for indicating the horizon and the elevation angles. To simplify the calculations I assume the earth to be spherical and smooth.

"The accompanying diagram [*top of page 128*] depicts the earth as it might appear in cross section if it were cut through the meridians 80°W and 100°E. Note that a straight line tangent to the earth's surface at 40°N defines the horizon and is intercepted by the orbit of the satellite at distances from the observer of 35.6 degrees of arc. The points of interception mark the limits beyond which the receiving station could not pick up line-of-sight signals from a weather satellite that orbits at a height (h) of 1,464 kilometers (910 miles) above the earth's surface. The diagram is based on the U.S. weather satellite NOAA-2, which derives its acronym from the National Oceanic and Atmospheric Administration.

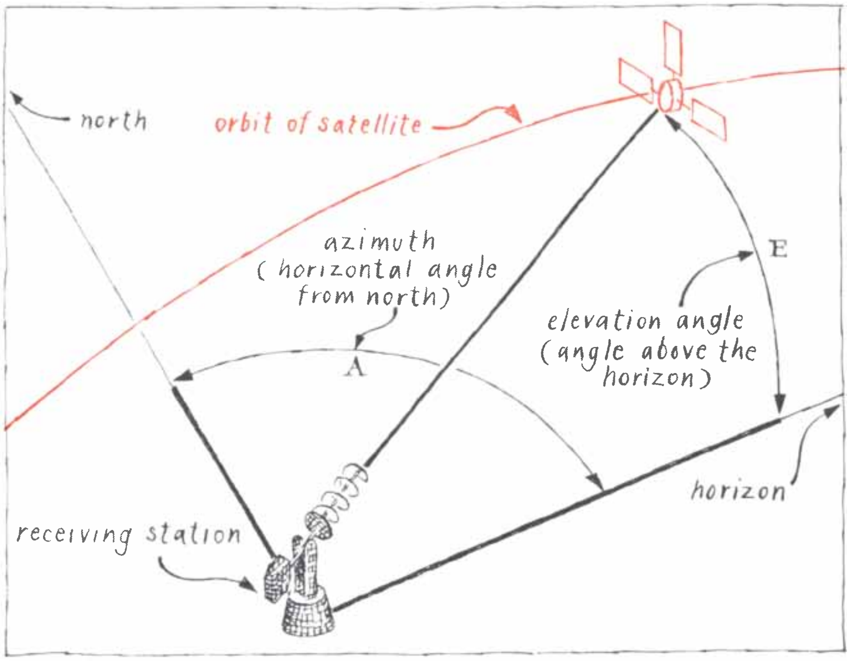
"Observe that an oblique triangle can be drawn by connecting with straight lines the center of the earth, the observing station and the satellite. In this triangle the lengths of two sides are known. One side is the earth's radius, R. In the calculations that follow R is assigned the value of 6,378 kilometers. The other known side is equal to the sum of $R + h$, 7,842 kilometers. The largest angle of the triangle, the one that faces the large known side, is also known. This angle is the sum of the assumed angle of elevation and the right angle made by the vertical line through the station and the tangent to the earth's surface.

"With this information one can easily determine the size of the angle (B) between the known sides and therefore the length of the arc at the earth's surface that is subtended by the known sides. This arc comprises the radius of the circle of points that could lie directly under the satellite when the satellite is at the assumed angle of elevation above the horizon. For example, the triangle that has been drawn with the darkest lines in the diagram corresponds to an elevation angle of 20 degrees.

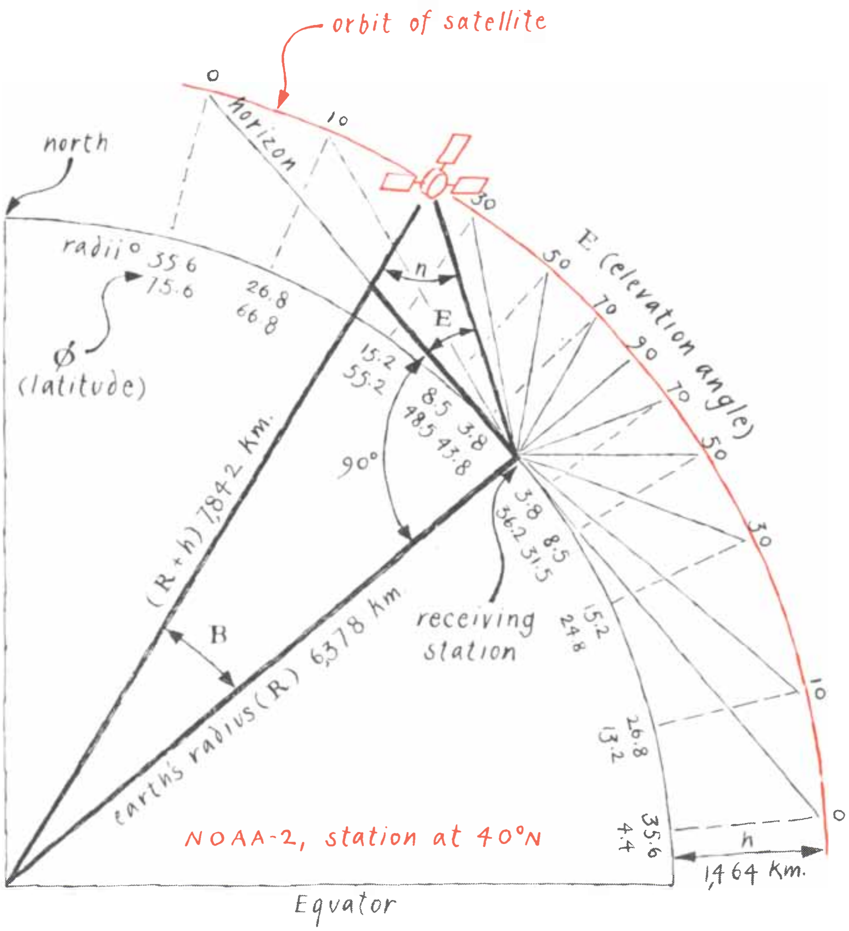
"The sequence of arithmetical opera-



William K. Widger's device for plotting a satellite's orbit



Angles related to a position of a satellite



Geometry of elevation circles

tions required to determine the corresponding radius (B) in degrees of arc for a circle of subpoints at 20-degree elevation is specified by the first of the accompanying formulas [*top of page 130*]. Divide the earth's radius by the sum of the radius and the height of the satellite: 6,378 / (6,378 + 1,464) = .8133. From a table of trigonometric functions determine the cosine of each desired angle of elevation. In this example the desired elevation is 20 degrees. According to the table, cos 20° = .9397. The formula requires the product of $[R / (R + h)] \times \cos E$, which in this example is .8133 × .9397 = .7642.

"Next the formula requires the angle (arc cosine) that corresponds to this cosine. According to the trigonometric table, the cosine .7642 corresponds to the angle 40.2 degrees. Finally, the formula states that the desired radius (B) in degrees of arc is found by subtracting the angle E (20 degrees) from the arc cosine angle: $B = 40.2 - 20 = 20.2°$.

"To draw on the map a circle that represents points in all directions that are 20 degrees above the local horizon, place the point of a pen compass on the equator of the graph paper at 0 degrees longitude and open the instrument until the nib of the pen corresponds to latitude 20.2 degrees. With this distance as the radius, transfer the point of the compass to the location of the observing station on the map and draw the circle. Similarly determine the radii and inscribe circles at other angles of elevation as desired. In the case of NOAA-2 I have tabulated radii for elevation angles at intervals of 10 degrees from 0 to 90 degrees [*see bottom illustration on page 130*].

"Place the clear plastic overlay over the map and pin it at the Pole so that the overlay can be rotated around the pin. The path of the satellite will be plotted on the overlay in terms of the object's changing latitude and longitude. These coordinates can be calculated for a track of any length in increments of, say, 10 arc degrees, beginning on the Equator at 0 degrees longitude. Each length will be the hypotenuse of a right triangle as drawn on the surface of a sphere [*see illustration on opposite page*]. The lengths of the corresponding legs of the triangle are measured in degrees of latitude and longitude.

"The fact that members of the solar system rotate in step somewhat like the meshed gears of a machine enables one to plot the course of an earth satellite with a surprisingly small amount of initial information. For example, all that is needed to plot the course of NOAA-2 is the fact that it travels 1,464 kilometers above the earth's surface in a sun-synchronous orbit. The plane of a sun-synchronous orbit turns one full revolution with respect to the sun in a year of 365.2422 days. The angle made between the orbital plane and the sun remains constant. Moreover, the ascending node (the point at which a northward-bound earth satellite crosses the plane of the celestial equator—not, as in the usual case, the plane of the ecliptic) occurs at the same local time on each orbit.

"Offhand one might suppose that the orbit of any earth satellite would remain stationary in relation to the fixed stars. It probably would if the universe were symmetrical. Actually the orbits of earth satellites are strongly influenced by the aspherical shape of the earth, which bulges slightly at the Equator. Satellites that orbit in planes making an angle other than 0 or 90 degrees with respect to the earth's Equator respond to the gravitational pull of the bulge much as a gyroscope behaves when it is supported at one end of its shaft. Instead of falling the gyroscope rotates around the point of its support. In the same way the gravitational force between the bulge of the earth and the satellite causes the orbit of the satellite to precess in relation to the fixed stars. This effect is observed in the wobbling motion of an inclined top. Indeed, the rate at which the orbital plane of an earth satellite rotates varies with both the altitude of orbit and the angle at which the plane of the orbit is inclined with respect to the Equator.

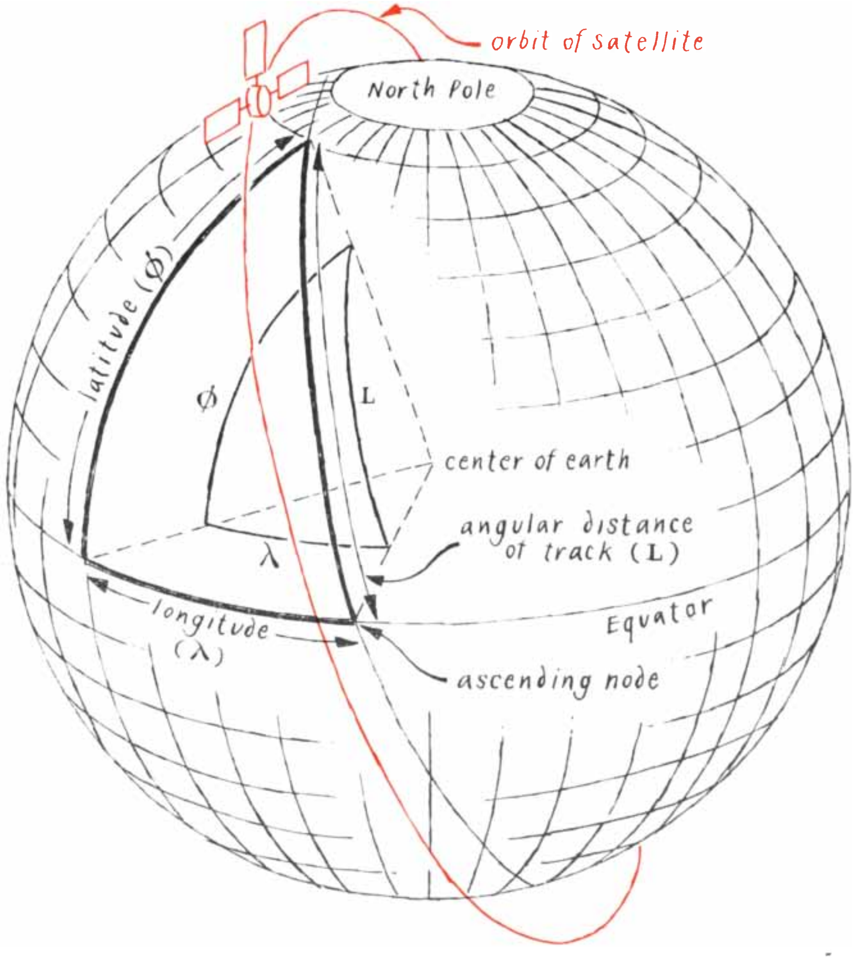
"To establish the sun-synchronous condition, weather satellites such as NOAA-2 are injected into orbit at an extremely precise angle of inclination, which is based on mathematical relations worked out by specialists in astrodynamics. If the height of the orbit is known, the sun-synchronous angle of inclination (i) can be calculated, as shown in Formulas 2 and 3 [*top of next page*]. In the case of NOAA-2 the inclination of the orbit (i) is equal to the angle (arc cosine) that corresponds to the cosine of $-4.7349 \times 10^{-15} \times (R + h)^{7/2}$. To find the 7/2 power of $R + h$ for NOAA-2, multiply by itself seven times the square root of the sum of 6,378 + 1,464. The square root of the sum is 88.56, and 88.56^7 is 4.2723×10^{13}. The product of $4.2723 \times 10^{13} \times -4.7349 \times 10^{-15}$ is $-.2022$.

"The angle that is equivalent to the negative cosine $-.2022$ is found by subtracting from 180 degrees the angle that corresponds to (positive) cosine .2022: arc cos $.2022 = 78.33$; arc cos $-.2022 = 180° - 78.33° = 101.67° = i$, which is the desired angle of inclination of the satellite NOAA-2. Formulas that determine the track (L) on the surface of the earth over which satellites such as NOAA-2 orbit make use of the supplement (I) of the angle of inclination: $180° - 101.67° = I$.

"Also required to calculate the track (L), as we have seen, is the period of time in minutes (P) necessary for the satellite to complete one orbit of the earth beginning and ending at the ascending node. The period (P), as indicated by Formula 4, is equal to 1.6586×10^{-4} multiplied by the cube of the square root of $R + h$. For NOAA-2 the square root of $R + h$ is 88.56. The cube of 88.56 is 6.9456×10^5. Therefore the nodal period of NOAA-2 (P) is $1.6586 \times 10^{-4} \times 6.9456 \times 10^5 = 115.19$ minutes.

"With this information one can calculate intervals of latitude (ϕ) and longitude (λ) that comprise the legs of the spherical triangle that has as its hypotenuse the track (L). When I design a prediction device, I normally begin by assuming that the track starts on the Equator at 0 degrees longitude. I then compute the corresponding latitude and longitude as the satellite advances along its track in increments of 10 degrees of arc. A somewhat more accurate curve could be plotted by reducing the increments to five degrees of arc, which would double the number of computations.

"Latitude (ϕ) is determined with the aid of Formula 5. For example, when NOAA-2 has orbited through a distance of 10 degrees of arc from its assumed starting point above the Equator, it arrives at the latitude: arc sin (sin 78.33° $\times$ sin 10°) $= .9793 \times .1736 = .17 = 9.8° = \phi$. After NOAA-2 has advanced 90 arc degrees along its track its latitude has increased to arc sin (sin 78.33° $\times$ sin 90°) $= .9793 \times 1 = .9793 = 78.33°$, its closest approach in latitude to the North Pole. When NOAA-2 completes half of its orbit (180 degrees), its latitude falls to arc sin (sin 78.33° $\times$ sin 180°) $= .9793 \times 0 = 0°$. Stated differ-



Orbit of a satellite as defined by a spherical triangle

B	$= \arccos [\{R/(R+h)\} \cos E] - E$	(1)
i	$= \arccos [-4.7349 \times 10^{-15} (R+h)^{7/2}]$	(2)
I	$= 180° - i$	(3)
P	$= 1.6586 \times 10^{-4} (R+h)^{3/2}$	(4)
θ	$= \arcsin (\sin I \sin L)$	(5)
λ	$= \{ \arctan (\cos I \tan L) \} + PL / 1{,}440$	(6)
θ_t	$= \arcsin [\sin I \sin (360t / P)]$	(7)

Formulas for predicting an orbit

ently, the satellite has arrived above the Equator on the opposite side of the earth. From 180 to 360 degrees the trigonometric sign of all angles is negative, indicating that these latitudes lie in the Southern Hemisphere.

"Each increment of longitude (λ) constitutes the second leg of the associated spherical triangle. The increments are calculated with the aid of Formula 6. Longitude is generated by two motions: the westward advance of the satellite on its retrograde orbit and the slower motion of the observer as the earth's rotation carries the station eastward. The component of the longitudinal motion for which the satellite is responsible is described by the formula as an arc that corresponds in degrees (arc tangent) to a trigonometric tangent expressed as the product of cosine *I* multiplied by tangent *L*. To this arc, according to the formula, must be added the second component of motion: an arc equal in degrees to the product of $P \times L / 1{,}440$.

"Assume that NOAA-2 has advanced 10 degrees along the orbit (along the hypotenuse of the spherical triangle). It has reached the latitude 9.8°N, as has been previously calculated. Simultaneously the satellite has moved westward along the inclined orbit: $L = 10°$. As has been previously determined, the cosine of *I* (78.33 degrees) is .2022. The trigonometric table lists the tangent of 10 degrees as .1763. Substitute these values in the formula: arc tan $(.2022 \times .1763) =$ arc tan $.0356 = 2°$. This is the arc through which NOAA-2 would have orbited in longitude if the earth were not rotating.

"The motion of the earth generated an interval of longitude: $P \times L / 1{,}440 = 115.19 \times 10° / 1{,}440 = .8°$. The total increase in longitude is therefore $\lambda = 2 + .8° = 2.8°$. Near the Equator, where the meridians are widely spaced, the rate at which the satellite moves westward is low in relation to its motion northward. Over the first 10 degrees of NOAA-2's orbit from the Equator its longitude increased by only 2.8 degrees, an average of only .28 degree of longitude per degree along its track.

"This rate increases dramatically as the satellite approaches the poles, where the meridians meet. When *L* is 90 degrees, the tangent of *L* is infinity. The product of cosine *I* times infinity is infinity. The arc tangent of infinity is 90 degrees. At this point along the track the component of longitude generated by the earth's rotation is $115.19° \times 90 / 1{,}440 = 7.2°$. Therefore $\lambda = 90° + 7.2° = 97.2°$. When *L* has increased to 100 degrees, the equation is $\tan L = 180° - 100° = \tan 80°$. According to the laws of trigonometry, tangents of angles that lie between 90 and 180 degrees are negative numbers. Hence $\tan 100° = -\tan 80° = -5.6713$. Substitute the appropriate values in the formula, cosine *I* and tangent *L*: $.2022 \times -5.6713 = -1.1467$. Therefore at this point along the track the component of arc for which the satellite is responsible is $180° +$ arc tan $-1.1467 = 180° - 48.9° = 131.1°$.

"To this longitude must be added the increment contributed by the earth's rotation: $115.19 \times 100° / 1{,}440 = 8°$. Hence after NOAA-2 has orbited 100 degrees from the Equator its longitude has increased to $\lambda = 131.1° + 8° = 139.1°$. During the 10 degrees of arc along the track from 90 to 100 degrees the longitude of the satellite shifted westward from 97.2 to 139.1 degrees, an advance of 41.9 degrees compared with the difference of only 2.8 degrees for the same length of track at the Equator!

"When you do the arithmetic, keep in mind that the tangents of angles that lie between 180 and 270 degrees are positive. The calculation of these angles is similar to that of angles between 0 and 90 degrees. The accompanying table lists the latitudes (ϕ) and longitudes (λ) that determine points at each 10 degrees of arc along the track of NOAA-2. Beginning on the Equator, plot the points on the plastic overlay. (The overlay can be anchored temporarily at the edge of the map with bits of adhesive tape.) Connect the points by a graph to represent the track. The graph can be drawn smoothly in india ink with a ruling pen and a draftsman's French curve.

"The track can be calibrated in minutes of time since the satellite crossed the ascending node. Intervals of two minutes are convenient. I prefer to plot the intervals in terms of latitude (ϕ_t) with the aid of Formula 7. For example, two minutes after NOAA-2 crosses the Equator its latitude, according to the formula, is arc sin [sin $78.33 \times$ sin $(360 \times 2 / 115.19)] =$ arc sin $.1066 = 6.1° = \phi_t$. If *P* is known in minutes, the formula can be simplified. Divide 360 by the known period. Thereafter multiply the intervals of time (*t*) by the resulting constant. For NOAA-2 the constant is $360 / 115.19 = 3.125$. The formula becomes arc sin (sin *I* sin 3.125 *t*). The accompanying table [*bottom of this page*] lists latitudes for calibrating the track of NOAA-2 during the first hour of its orbit from the ascending node.

"If two or more tick marks are inscribed on the overlay at the Equator, the device will serve for predicting the location of the satellite for as many as about 12½ orbits. The marks are spaced at intervals of longitude equal to the satellite's apparent westward movement during each full orbit, as measured from consecutive ascending nodes. This distance, in degrees of arc, is about .25 times *P*. For NOAA-2 the tick marks are spaced at intervals of $.25 \times 115.19 = 28.8°$. In general predictions of more than 24 hours are not useful because small errors build up rapidly.

"To use the instrument rotate the overlay to the point at which the zero time of the track coincides with the longitude of the satellite at its next ascend-

ELEVATION CIRCLES		SATELLITE TRACK DATA						TIME FROM ASCENDING NODE					
E	B	L	θ	λ	L	θ	λ	t	θ_t	t	θ_t	t	θ_t
0	35.6	−20	−19.6	−5.8	100	74.7	139.1	−6	−18.3	18	54.5	42	47.4
10	26.8	−10	−9.8	−2.8	110	67.0	159.7	−4	−12.2	20	60.3	44	41.4
20	20.2	0	0	0	120	58.0	170.3	−2	−6.1	22	65.9	46	35.4
30	15.2	10	9.8	2.8	130	48.6	176.8	0	0	24	71.1	48	29.3
40	11.5	20	19.6	5.8	140	39.0	181.6	2	6.1	26	75.5	50	23.2
50	8.5	30	29.3	9.1	150	29.3	185.3	4	12.2	28	78.1	52	17.1
60	6.0	40	39.0	12.8	160	19.6	188.6	6	18.3	30	77.7	54	11.0
70	3.8	50	48.6	17.6	170	9.8	191.6	8	24.4	32	74.7	56	4.9
80	1.9	60	58.0	24.1	180	0	194.4	10	30.5	34	70.1	58	−1.2
90	0	70	67.0	34.7	190	−9.8	197.2	12	36.6	36	64.8	60	−7.4
		80	74.7	55.3	200	−19.6	200.2	14	42.6	38	59.2	62	−13.5
		90	78.3	97.2	210	−29.3	203.3	16	48.6	40	53.3	64	−19.6

Data for a plotting device designed for the satellite NOAA-2 ing node near the longitude of the observer. This information is broadcast worldwide by radio station W1AW, as Eugene F. Ruperto explains in his description of an amateur weather-satellite station in this department [SCIENTIFIC AMERICAN, January].

"The fact that the orbit of the satellite can penetrate the observer's horizon as plotted on the map, or even some of the inner circles of elevation, does not necessarily ensure that radio signals from the satellite can be picked up. The device is designed on the unrealistic assumption that the earth is a smooth sphere. Local features of the terrain may interfere with the line-of-sight path. Experience at a given location will soon determine for the observer the bands of ascending-node longitudes and the corresponding interval of time after the ascending node when the satellite is within receiving range of the station.

"Incidentally, lest readers misinterpret Ruperto's discussion of the information picked up and broadcast by NOAA-2, the scan to which he refers lies only across the earth at a right angle with respect to the satellite's track. The scans are analogous to the horizontal lines of a television picture. The signal in the infrared scan is the sum of all energy in the infrared band from 10.2 to 12.5 micrometers (not nanometers). The area resolved in the infrared is about four miles across when the scan is directly under the satellite. The visible channel integrates as a single signal all reflected solar energy between .5 and .7 micrometer, with a resolution of about two miles when the radiometer is looking straight down. At a distance of 500 miles on either side of the subpoint these resolutions are degraded by a factor of approximately 1.5.

"Amateurs should keep in mind the fact that although the visible channel picks up a true picture of cloud cover, the infrared channel broadcasts a temperature map of the earth's surface and of the highest clouds except thin cirrus clouds. Low temperatures appear as shades ranging from white to light gray, higher temperatures are in shades from dark gray to black.

"The tops of clouds at high and intermediate altitudes are cold. They appear white. For this reason the images of cloud cover made in infrared, even when they are picked up on the dark side of the earth, frequently resemble pictures made in visible light. Thin cirrus clouds and clouds with very low (warm) tops may not be visible in infrared against the background of the earth's surface."

BOOKS

The functions of sleep, microbes old and new and the art of workmanship

by Philip Morrison

THE FUNCTIONS OF SLEEP, by Ernest L. Hartmann, M.D. Yale University Press ($8.50). Why do we sleep? In a personal, clear (although rather medical) prose this Tufts professor of psychiatry, a prolific researcher on sleep, dares to propose in this paperback an answer to the general reader and the specialist alike. He is a man undaunted by anyone's technique or jargon. He asserts the joint utility of concepts at very different stages of abstraction, citing with sober appraisal studies of the level of brain tyrosine hydroxylase no more or less knowingly than he refers to ego regression. Let the origin of the ideas lie in "electrical studies of a giant squid axon" or in "hours spent listening to a neurotic Viennese patient"; the new psychology will arise from both. His demanding essay is a beginning toward a grand goal ahead, a powerful theory of the relationship of mind and body.

The laboratory has provided two main lines of inquiry into the human sleeper. One is electroencephalography: the valuable if frustrating recording of those submillivolt waves at five or 10 cycles per second, an effort to study the big computer inside by measuring leakage currents in its front panel. The second is human response to drugs, viewed with a lively sense of analogy to the neurochemistry of the rat. Add extensive clinical tests of deprivation and stress, questionnaires and interviews with long and short sleepers and other such "experiments in nature" and his evidence is before you, often graphically. The literature is voluminous, since clinical workers are many. The 300 and more entries in the bibliography are a mere sample; several thousand papers have appeared since 1966, when the author wrote his first review of the field.

Sleep is not an EEG tracing, of course, but a recurrent state of behavior marked by a much dulled response to outside stimuli and a certain quiescence. The EEG and the polygraph have shown us particular electrical states associated generally—not always—with that behavior. Our sleep is cyclical, a cycle lasting from an hour and a half to under two hours. Each sleeper spends about three-fourths of each cycle in S sleep (the S standing for the synchronized delta waves, with from one to four peaks per second, that mark the deepest stages of this form) and the rest of the cycle in "paradoxical" *D* sleep: desynchronized or dreaming sleep, with rapid eye movements, irregular pulse, penile erections and an EEG pattern resembling the lightest sleep, and yet with minimal large-muscle potentials and maximal thresholds of arousal. We share the types of sleep with most mammals (perhaps with all but the spiny anteater), with birds (which show a very little *D* sleep), maybe even with reptiles, but not with fish. *D* sleep is neurologically active, with some single nerve cells as busy as when we are awake. It is striking that young mammals sleep the most, with a greater share of *D* sleep; there are some signs that the protein-synthesis rate rises during S sleep, but hardly any chemical differences between sleep and waking are clear.

Prolonged sleeplessness kills; it takes a couple of weeks in dogs, rats and cats. It is hard to see real changes in the abilities of sleep-deprived humans, except that they tend to fall asleep during the tests! This result is disappointing, but Professor Hartmann argues that the "obvious subjective and emotional effects" of the deprivation of sleep leave the impression of damage to the focused attention. "Dream deprivation" is a fascinating goal, and it is now possible to reduce the *D* time specifically—although the brain cleverly rearranges its patterns, transferring the important diagnostic bursting spikes from special parts of the brain to S periods or even to waking. We are inveterate dreamers, it seems, who find our way to dreamland along many paths.

Advertise in the "daily newspapers in Boston and New York" and you can draw a couple of hundred adult males who always sleep more than nine hours or less than six hours each day. These men filled out sleep logs (for pay) and answered many questions; after further selection 29 came to the laboratory for eight nights of sleep study. Long sleepers spend nearly twice as much time in *D* sleep as short sleepers, although about proportionately the same amount of time. Psychologically, long sleepers are worriers, critical, uncomfortable; short sleepers are men who tend to "shrug things off and not get involved." Another couple of thousand people were questioned and interviewed in a search for sleepers whose duration of sleep varied with circumstances. The data seem to be mere impressions, but they are strong. Soldiers warned of impending combat sleep long hours; young women clearly relate times of acutely increased sleep to periods of emotional loss and grief; one man whose life changed from that of a laborer to that of an anxious university student (in a special program for intelligent people who had not attended college) found he needed one or two more hours of sleep per night. It has proved hard to induce such variation experimentally. Young people wearing inverting prisms—seeing their world upside down—need much *D* time at first, but the data are not beyond question. A group required to spend four hours each evening in a "perplexing atmosphere" (performing difficult tasks without explanation) needed more *D* time.

The pattern is there even if it is not yet sharp: stress, worry, "reprogramming" and the like are correlated with, and sometimes produce, increased *D* time and more sleep. In the laboratory electroshock and certain antidepressant drugs (the monoamine oxidase inhibitors) can sharply reduce *D* time. More impressive still, *D* time is increased by drugs that reduce the amount of the essential transmitter substance norepinephrine at the synapses; treatments that reduce *D* time make more norepinephrine available.

Simple tiredness also takes two forms. Here Professor Hartmann is persuasively clinical, even psychoanalytical. No graphs now but a long table making his case that when we become physically tired, without much anxiety, we seek sleep easily; when emotional stress and tension leave us "mentally" tired, we attain the sleep we need more slowly and are meanwhile headachy, irritable, unable to plan well or adapt smoothly. The two states may represent two distinct needs, the first for the more general and physically restorative S sleep, the second for specialized repair of the subtle feedback processes that can be described as ego functions.

Dreams themselves, treated here very sketchily but seriously, can be embraced in this model. The most obvious aspect of dreams is perhaps the easy acceptance of the bizarre and the unusual. An entire function is shut down: the judgment of reality. By now the author's exciting if tentative conclusion is almost easy to guess. S sleep is physically restorative; it differs from fully relaxed wakefulness mainly in specific brain electrophysiology. This suggests that it reflects a catabolic process within the brain; the hour-long time scale may imply that some relatively complicated biosynthesis of some large molecules is involved. *D* sleep, for which prior and longer S sleep seems a necessary precondition, serves rather to restore some special state of the neurotransmitters called catecholamines that are necessary for changes in synaptic linkage in particular regions of the cortex. *D* sleep is rather the time for rich reconnections according to some high-level computer language within those domains. The amphetamines also make a surplus of catecholamines available; they seem to do (or overdo) what *D* sleep does. Such drugs induce a state of overfocusing or "riveted attention"—in the end frank paranoia.

No one will take this somewhat wobbly pyramid of hypotheses for the final word, but we owe the courageous author a considerable debt for giving us this prefiguration of the "music of the future." Neither this or that -ism nor any single scheme of abstraction will lead to a coherent theory. The road is long. Professor Hartmann himself reminds us that his two functions of sleep resemble the analysis made long ago by a distraught Macbeth: "Sleep that knits up the ravell'd sleave of care... / Chief nourisher in life's feast."

THE BLUE-GREEN ALGAE, by G. E. Fogg, W. D. P. Stewart, P. Fay and A. E. Walsby. Academic Press ($24). AGENTS OF BACTERIAL DISEASE, by Albert S. Klainer and Irving Geis. Harper & Row, Publishers, Inc. ($9.95). First part the quick from the dead. Then divide the living into two groups. On the one hand there will be all the forms of life from amoeba to elephant, from kelp to sequoia. On the other there will be two and only two groups of living forms: the bacteria and the blue-green algae. "Blue-green algae are not always blue-green in colour but... one can recognize them instantly under the microscope because of the characteristically homogeneous appearance which the absence of... organelles... gives to their protoplasm." These smallest and simplest forms of life are the topics of two different volumes.

The Blue-Green Algae is the work of a lifelong student of this ancient strain and the younger colleagues working in his laboratory at Westfield College of the University of London. Their subject has undergone an exhilarating upsurge of interest, fed by the power of microbiology. Where once the blue-greens seemed obscure, slow-growing nuisances of nutrient-rich conditions, now they are seen in a truer light: ancient precursors of higher forms, probably the first donors of the very oxygen of our atmosphere, stubbornly prospering on the frigid lakes of the Antarctic, in the hot springs of Yellowstone, on damp walls, in arid deserts, in tidal flats and in rice paddies, all over the world except in Australia. Only bacteria range more widely.

Classical microbiological techniques make it possible to isolate a pure inoculum of blue-greens, given plenty of patience. Their mucilaginous sheaths harbor embedded bacteria that contaminate the cultures grievously. Once pure, some species of these algae will grow in a fully mineral medium, feeding on bubbling air and light from a fluorescent tube, their biomass doubling every two hours. They fix carbon photosynthetically from carbon dioxide as green plants do, and with the most important form of chlorophyll. They can fix nitrogen from the air; it seems probable that the half of mankind who live on rice are largely dependent on blue-green algae for the bounty of their unfertilized paddy.

The blue-green algae, like bacteria, occur as single cells or in clumps, necklaces and even branching filaments. Many species have a couple of types of differentiated cell. Certain filamentary forms produce chains of identical cells, but once in a while along the chain there is a special bead, thickly sheathed and usually larger, called the heterocyst. The electron microscope shows a complex ultrastructure in the heterocyst, and it was recognized in the past few years (by these authors) that the heterocyst

Our Green Lite℠* Service is so good we left Don Rickles speechless.

Green Lite* Service is our new system that helps you pick up your car faster and return it more conveniently. And we knew if there were any flaws, Don Rickles would be the guy to find them.

"You hotshots are asking for it," he snickered filling out the Green Lite Service application.

After we mailed Don his Green Lite Service identification card, he immediately called to put us to the test. He gave his identification number and added with a sneer, "You want my shoe size, too?"

"No sir," replied our reservationist. "We've got all the information we need stored in our computer."

So when Don arrived at our counter he found his rental agreement filled out and waiting.

"Not bad," he conceded. "But now I bet I turn into a pumpkin waiting for my car."

"No sir," said our counter representative. "Just show me your driver's license, authorized credit card and sign."

"That's it?" asked Rickles.

"That's it," she answered. "And if you're in a hurry when you return the Buick Century we're renting you, use our express check-in box.† And of course, at most locations you still get our 26 hour check-in day, and you get S&H Green Stamps on all U.S. rentals.

"But... but..." stammered Rickles.

Our service is so good it left Rickles speechless. Pick up an application for Green Lite Service at any National counter.

NATIONAL CAR RENTAL

When it comes to service we mean business.

†Available at participating locations for use by charge customers. © National Car Rental System, Inc. 1974 (in Canada it's Tilden Rent-a-car; in Europe, Africa and the Middle East it's Europcar.) *Green Lite and the design of the semaphore are service marks of National Car Rental System, Inc.

cell is the seat of enzymatic fixing of the nearly inert nitrogen molecule. The nitrogen-fixing factory is set going by a deficiency of organic nitrogen in solution, but it is highly sensitive to poisoning by free oxygen. The nitrogen-fixing bacteria of the buried root nodules of legumes do not, of course, encounter much free oxygen, but the blue-greens release it themselves by photosynthesis; it seems plausible that the heterocyst originated in order to separate the two conflicting chemical processes, oxygen release and nitrogen reduction, both essential to the life of the organism. Some blue-greens without heterocysts can fix nitrogen, but only under extreme oxygen-poor conditions.

For 20 years silicified microfossils of cells that are pretty surely blue-greens have been pursued farther and farther back into the Precambrian eons. The layered colonial structures called stromatolites, the blue-greens' version of coral, date from as much as three billion years ago. It seems plausible that the blue-greens first invented oxygen release and then their heterocysts, and finally so polluted the atmosphere with the well-known microbial poison, oxygen, that the solar ultraviolet was itself dimmed by ozone; then larger forms developed, finally on dry land, feeding on the very poison the algae had excreted. Nitrogen-fixing was an algal trick too: they began to inherit the earth from the bacteria once abundant initial supplies of combined nitrogen had been used up. The case for retention of once free-living blue-greens as the photosynthetic organelles of green plant cells, now living inside the cells as obligate symbionts, is made with due prudence. These high matters may soon pass from conjecture to test in the microfossil record.

Blue-green algae move somehow by "a slow, smooth progression without there being a visible means of propulsion." They need a surface, and some trick of the mucilage layer is probably responsible. Two clever microphysiologists have measured the power involved, and they conclude that one gliding algal filament (speeding by at almost three microns per second) used about 7,300 molecules of ATP per second to move, less than one part in 2,000 of the overall respiration.

These botanists (the blue-greens seem less ambiguously plants than other microforms do) have prepared an up-to-date and thorough volume, perhaps the best overall treatment of this fascinating class of organisms to be found. It is well illustrated with electron micrographs and photographs and graphs of every kind, including color pictures of blue-greens in their varied habitats—one example even deliciously and properly blue-green. A watercolor plate by the senior author adds a sense of tradition to this modern, compendious book.

Agents of Bacterial Disease is another excellent survey of an entire range of life. Here, however, those widespread forms of bacteria that once dwelled in the earth and still cover it are not to be found; as the title makes clear, the emphasis is wholly on the more specialized bacteria that are true parasites on one primate species, our own dear selves. It was of course medical concern and effort that made bacteria the best known of microorganisms, no less than their large and swiftly grown populations and the easy chemical selection, the simple properties that first attracted the attention of the pioneers of modern molecular biology.

This book is a concise introductory text in medical bacteriology. It is unique in its emphasis on visual presentation. The 115 illustrations are particularly rich in scanning electron micrographs. Their depth of field gives a furry reality not only to the diagnostic forms of clinically important organisms but also to such processes as antibiotic exposure, which allow some bacteria to grow but not to divide, and thus to yield grossly elongated cells or to puff out long pouches of cell wall (a kind of exaggerated bacterial hernia) at the place of normal division. Chains of cells are seen linked by thin bridges of wall. One remarkable set of four pictures shows the same cells stained in an ordinary photomicrograph and in the phase microscope, transmission electron microscope and scanning electron microscope. It is a capsule history of technique in a couple of pages.

The book opens with a few introductory chapters, covering very briefly the biology of bacteria and immunology, and then devotes about a dozen chapters to various types of bacteria and the diseases they cause—as streptococci, clostridia and so on. The summary text describes the various tests, bacterial products, disease signs and symptoms and even treatments. Many tables, diagrams and drawings are included, the better to carry out the theme of visual presentation. The logical and explicit drawings are the work of the junior author, a well-known scientific illustrator whose work has often been seen in this magazine.

The Nature and Art of Workmanship, by David Pye. Van Nostrand Reinhold Company ($2.75). This little paperback, only 60 pages of text and another 30 of carefully captioned photographs, is the work of the Professor of Furniture Design at the Royal College of Art in London. His eye is naturally out for the quality of work in wood, and half of the photographs are of wood objects, but the rest span the range from sail canvas to beer can and computer module. It would be an error to try to enlarge much on the message of a book this crisp and thoughtful, but some remarks are called for. Altogether it is as penetrating an account of the aesthetics of what humans make as has been seen for a good many years, and it deserves a wide readership.

Such work rests on the validity of nice distinctions. Professor Pye begins with the sharp comment that design can be symbolically conveyed, whereas workmanship cannot. "On the workman's decision depends a great part of the quality of our environment." The designer prepares the score, the workman performs the music. Workmanship is of two kinds: the workmanship of risk and the workmanship of certainty. In the first the result is not predetermined but depends heavily on the maker; the quality of the second kind is fixed in advance, by jig or tool bed, die or punch. To separate hand and power tools is no useful way of division and to distinguish mass production from one by one is also a false division. Indeed, little is really done "by hand"; how can we exclude from the machine category the brace and bit, the potter's wheel and the hand loom? Writing with a pen is perhaps the commonest handwork that remains.

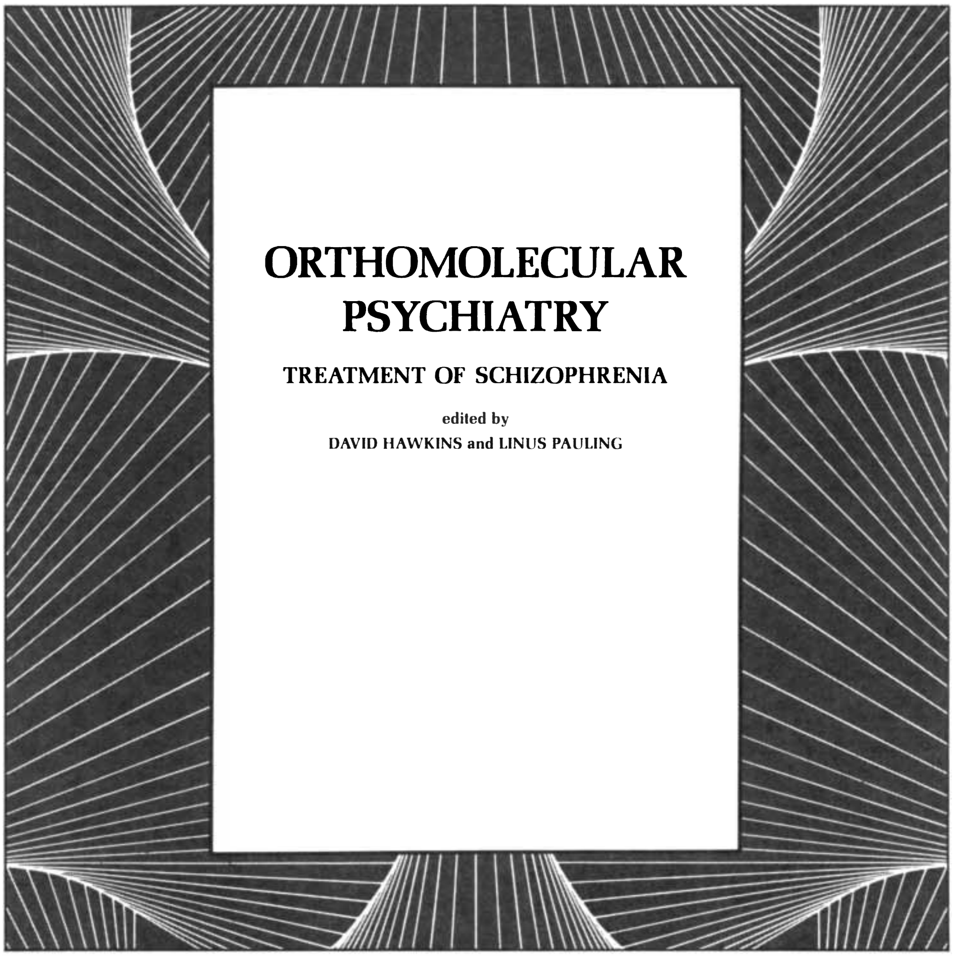
Nor can "rough" and "precise" be helpful dividers; rough work is often excellent, fully achieved, beautiful. Pye uses instead "regulated" and "free." Regulated work admits of no disparity between idea and achievement; when there are evident disparities, the work is called free. Both modes, risk and certainty, can, of course, achieve high regulation. The workman regulates risk by his dexterity, his gradualness and the use of jigs and forms. The workman of certainty usually gains regulation immediately; in this kind of production high regulation is today almost inevitable, although it is not logically required. So far the distinctions named are on the side of production. They seem helpful and knowing.

The key to Professor Pye's argument lies on the side of the user, the viewer, the perceiver. It is based on what he calls diversity. The intrinsic limitations of human vision and touch leave—under any circumstances of distance, lighting or size—some unresolved and indistin-

The world's first ambient light liquid crystal memory pocket calculator—the DataKing 800 manufactured by Rockwell International, can operate for one year on the same set of disposable batteries.

"The pieces of a vast puzzle are coming together. If they can be successfully assembled, they will constitute a medical revolution . . . *orthomolecular medicine*."

—Human Behavior



From Walter C. Alvarez, M.D.:
"Many thanks for sending me the magnificent book by Drs. Hawkins and Linus Pauling. I am delighted to see it because I am a believer in the work of Drs. Hoffer and Osmond, who years ago started this idea of treatment for schizophrenia. What has bothered me for years has been the tendency of nearly all physicians to say that the teatment is no good, when they have never tried it on patients.

"Now I hope that more doctors will try it."

From *Human Behavior* [May 1973] :
"The overwhelming weight of evidence seems to point in the same direction: Orthomolecular medicine is rational, and it works. . . .[This book is] a huge collection of research articles and clinical reports concerning this biochemical approach to mental illness. . . .fascinating. . . . this is all the stuff of scientific discovery."

"I join my coeditor, Dr. Hawkins, and the other contributors to this book in expressing the hope that it will be found useful not only by scientists and physicians but also by those who suffer from schizophrenia and by their families." —Linus Pauling

ORTHOMOLECULAR PSYCHIATRY: Treatment of Schizophrenia
Edited by DAVID HAWKINS, The North Nassau Mental Health Center, and
LINUS PAULING, Stanford University
1973, 697 pages, 92 illustrations, 79 tables, $17.00

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104
58 Kings Road, Reading, England RG1 3AA

guishable elements of any object. It is these that generally determine surface—through texture, reflections, gloss, luster and the rest. A surface looks tinny or leathery because we recognize by experience the particular scale and range of the unseen tiny undulations and deviations from flatness of the surfaces. The workmanship of risk can deal masterfully with these matters; "surface quality in man-made things comes of workmanship." Even a careful microfinish specification of surface roughness brings the designer little visual gain; he cannot hope to specify an entire surface, as complex as the Alps, with a number or two or even a root-mean-square average. Meanwhile the designer for the workmanship of certainty must continue to learn about the quality and potential of materials from the best works made by the craftsmen of risk.

"Four things are going wrong." One is the present inability of the workmanship of certainty to predetermine the diversity that makes surface; only in woven synthetic fibers and in the use of transparent and translucent plastics is there so far any sign of design success in diversity. Computer-controlled machine tools might begin to introduce regulated diversity but they have not done so yet. Another is the bad workmanship of risk in the assembly of objects made with high regulation, as in carelessly fitted millwork doors. Still another: The best of the regulated workmanship of risk requires so much time that it is dying of its high cost, although it still contributes unique qualities to our environment, for example the finest cabinetmaking. Finally, free workmanship is dwindling too (if less speedily).

The last three problems can be met by specific changes—through education, cooperation, deliberate support, or perhaps by amateurs who will work more for love than for money. The handmade gun and the handmade viola bow are examples of the finest regulated workmanship of risk today, and they are not in danger of neglect. Free workmanship too can be preserved; the Japanese packaging crafts give a splendid example. It is up to high technology and perceptual science, however, to solve the first problem. These are no matters of bread and butter, but we well know we do not live by bread alone.

No one who reads this book will again throw away a beer can with indifference. The slight buckling of the surface gives it an element of free workmanship, and the raw, direct edges do not clash with its low surface polish. "Anyone accustomed to doing regulated work by the workmanship of risk must feel something of a pang at throwing such a thing away, for to make it by the workmanship of risk would be an intensely difficult and very long job."

CHEMICALS FROM PETROLEUM: AN INTRODUCTORY SURVEY, by A. Lawrence Waddams. John Wiley & Sons ($6.95). First issued by Shell Chemical in Britain a dozen years ago, this expert and easily read book has become a standard training manual in the petrochemical industry. It conveys the feeling of an international industry as volatile as its feedstocks and products, and an economics as recycled and complex as its piping runs. The industry is the scene of ingenuity and analysis at every level, from the replacement of high-pressure reciprocating machines (mechanical wonders only a generation ago) by fast centrifugal compressors to the realization that the enticing economies of increased plant scale (cost varying with capacity to the .6 power) begin to pall once too many eggs are in one basket: the risk of start-up delay or plant malfunction grows at least in proportion to plant size. Although the book is up to date—three-fourths have been rewritten since the second edition only five years ago—it makes no reference to the uncertainties of crude-oil supply that are now so much on the public mind.

The story of the industry is told well, with a brief background in history, a long set of detailed chapters on the main chemicals (arranged by molecular size), a few chapters on economic and consumption matters and useful flow diagrams and photographs of plant.

The burden of the account is variations on a theme: engineers can easily handle liquids and gases on a large scale but not solids. Just as our fuel—several tons per year per human head worldwide—has tended to become fluid, so have the raw materials of chemical production become fluids, in part the direct by-products of fluid-fuel production. The most wanted products, however, are solid at the point of use (with the important exception of sulfuric acid and of ammonia solutions used directly for fertilizer). Nearly half of all elemental sulfur is now recovered from petroleum or "sour" natural gas.

Take tires. The commonplace automobile tire is about a third carbon black and two-thirds butadiene-styrene polymer. Carbon black is made nowadays by burning in a special furnace a fine spray of a rather heavy petroleum feedstock. Photographs of old plants give eloquent witness to how unsavory a pall of black soot used to be tolerated by the rural consumers of the air above the gas fields of Texas; the new furnace-black plant appears spotless. Butadiene? In Europe it is a solvent-separated product of the stream of four-carbon molecules arising from the high-temperature, steam-aided cracking of a feedstock rather like gasoline. Styrene? It is made in two steps from the catalytic alkylation of benzene with ethylene. Benzene? Once the principal aromatic-ring product of coal tar from the great coking furnaces, it now comes from catalytic hydrogenation in the petroleum refinery. By now most U.S. benzene is petroleum-born (more than 90 percent). In Europe too: in 1970 three tons of benzene came from oil for every one made from coal. Coke is still essential for steel and coal-gas manufacture, and benzene is still a side product. Those industries have been static, however, compared with the rising demand for petroleum aromatics. The shift from coal to oil-derived benzene took 15 years. Ethylene? This double-bonded two-carbon molecule is produced in the largest volume of any pure organic compound (ammonia is inorganic by current convention). In the U.S. it has been produced mainly by the pyrolysis of our ethane-rich natural gas; in Europe it is made by cracking a liquid petroleum fraction, since the natural gas of Europe is almost pure methane.

Tires, the various plastics—all the vinyls, polyethylene, polystyrene, polypropylene, phenolics, nylon, acrylics—ammonia, fertilizer and sulfur itself now derive from air and the automated and continuous flow of the petroleum industry. How closely the economics and the chemistry are linked to the much larger production volume of liquid and gaseous fuel is an important theme of the story; how that will resist change is unclear.

The economics of petroleum will not remain long in its present disarray but it is sure that costs will rise. The era of the chemical engineer's delight, with raw materials piped and tanked to the plant in a smooth fluid flow at the lowest prices in an ever increasing stream, is nearing an end. All curves of growth flatten in time. Solid coal will come back and itself be turned into fluid feedstocks. A tenth or so of the flood of fuel should meet all our needs for petrochemicals; perhaps it is true that "petroleum products are better used to make something tangible and useful, as an alternative to being burned to produce energy." Surely that will describe the industry better a generation hence, however useful our wrappings, roofing and gimcrack toys may be deemed to be in that wiser age.

INDEX OF ADVERTISERS

MAY 1974

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

NUCLEAR STRATEGY AND NUCLEAR WEAPONS

THE FUTURE OF THE SEA-BASED DETERRENT. Edited by Kosta Tsipis, Anne H. Cahn and Bernard T. Feld. The MIT Press, 1973.

CAN NUCLEAR DETERRENCE LAST OUT THE CENTURY? Fred Charles Iklé in *Foreign Affairs,* Vol. 51, No. 2, pages 267–285; January, 1973.

THE MUTUAL-HOSTAGE RELATIONSHIP BETWEEN AMERICA AND RUSSIA. Wolfgang K. H. Panofsky in *Foreign Affairs,* Vol. 52, No. 1, pages 109–118; October, 1973.

FLEXIBLE MADNESS? Herbert Scoville, Jr., in *Foreign Policy: 14,* pages 164–177; Spring, 1974.

THE ARCHAEOLOGY OF WINCHESTER

EXCAVATIONS NEAR WINCHESTER CATHEDRAL, 1961. Martin Biddle and R. N. Quirk in *The Archaeological Journal,* Vol. 119, pages 150–194; 1962.

EXCAVATIONS AT WINCHESTER: INTERIM REPORT Nos. 2–10. Martin Biddle in *The Antiquaries Journal,* Vol. 44, pages 188–219, 1964; Vol. 45, pages 230–264, 1965; Vol. 46, pages 308–332, 1966; Vol. 47, pages 251–279, 1967; Vol. 48, pages 250–284, 1968; Vol. 49, pages 295–329, 1969; Vol. 50, pages 277–326, 1970; Vol. 52, pages 93–131, 1972; Vol. 54, in press.

ARCHAEOLOGY AND THE HISTORY OF BRITISH TOWNS. Martin Biddle in *Antiquity,* Vol. 42, No. 166, pages 109–116; June, 1968.

LATE SAXON PLANNED TOWNS. Martin Biddle and David Hill in *The Antiquaries Journal,* Vol. 51, pages 70–85; 1971.

VISUAL PATHWAYS IN ALBINOS

ABERRANT VISUAL PROJECTIONS IN THE SIAMESE CAT. D. H. Hubel and T. N. Wiesel in *Journal of Physiology,* Vol. 218, pages 33–62; 1971.

A STUDY OF NORMAL AND CONGENITALLY ABNORMAL RETINOGENICULATE PROJECTIONS IN CATS. R. W. Guillery and J. H. Kaas in *The Journal of Comparative Neurology,* Vol. 143, No. 1, pages 73–100; September, 1971.

GENETIC MECHANISMS DETERMINING THE CENTRAL VISUAL PATHWAYS OF MICE. R. W. Guillery, G. L. Scott, B. M. Cattanach and M. S. Deol in *Science,* Vol. 179, pages 1014–1016; 1973.

CONGENITALLY ABNORMAL VISUAL PATHWAYS IN MINK (*MUSTELA VISON*) WITH REDUCED RETINAL PIGMENT. K. J. Sanderson, R. W. Guillery and R. M. Shackelford in *The Journal of Comparative Neurology,* Vol. 154, No. 3, pages 225–248; April 1, 1974.

THE TOP MILLIMETER OF THE OCEAN

ELECTRIFICATION OF THE ATMOSPHERE BY PARTICLES FROM BREAKING BUBBLES. D. C. Blanchard in *Progress in Oceanography: Vol. I,* edited by M. Sears. Pergamon Press, 1963.

JOURNAL OF GEOPHYSICAL RESEARCH. Vol. 77, No. 27; September 20, 1972.

GLYCOPROTEINS

GLYCOPROTEINS: THEIR BIOCHEMISTRY, BIOLOGY AND ROLE IN HUMAN DISEASE. Robert G. Spiro in *The New England Journal of Medicine,* Vol. 281, No. 18, pages 991–1001, October 30, 1969; No. 19, pages 1043–1056, November 6, 1969.

THE SYNTHESIS OF COMPLEX CARBOHYDRATES BY MULTIGLYCOSYLTRANSFERASE SYSTEMS AND THEIR POTENTIAL FUNCTION IN INTERCELLULAR ADHESION. Saul Roseman in *Chemistry and Physics of Lipids,* Vol. 5, No. 1, pages 270–297; October, 1970.

GLYCOPROTEINS OF BLOOD CELLS AND PLASMA. Edited by Graham A. Jamieson and Tibor J. Greenwalt. J. B. Lippincott, 1971.

THE BIOCHEMISTRY OF PLANT LECTINS (PHYTOHEMAGGLUTININS). Halina Lis and Nathan Sharon in *Annual Review of Biochemistry,* Vol. 42, pages 541–574; 1973.

THE PARTICLES OF WEAR

THE FRICTION AND LUBRICATION OF SOLIDS: VOLS. I AND II. Frank P. Bowden and D. Tabor. Oxford University Press, 1954 and 1964.

A METHOD FOR THE STUDY OF WEAR PARTICLES IN LUBRICATING OIL. W. W. Seifert and V. C. Westcott in *Wear,* Vol. 21, No. 1, pages 27–42; August, 1972.

IMPROVEMENT OF DESIGN AND MATERIALS BY FAILURE ANALYSIS AND THE PROGNOSTIC APPROACH TO RELIABILITY. D. Scott and A. I. Smith in *Institution of Mechanical Engineers Conference on Selection of Materials in Machine Design,* Preprint 358/73. Institution of Mechanical Engineers, London, 1973.

COMPUTER GRAPHICS IN ARCHITECTURE

SKETCHPAD III: A COMPUTER PROGRAM FOR DRAWING IN 3-DIMENSIONS. Timothy Johnson. Massachusetts Institute of Technology Electronic Systems Laboratory, June, 1963.

THE ARCHITECTURE MACHINE. Nicholas Negroponte. The MIT Press, 1970.

COMPUTER DISPLAYS. Ivan E. Sutherland in *Scientific American,* Vol. 222, No. 6, pages 56–81; June, 1970.

PRINCIPLES OF INTERACTIVE COMPUTER GRAPHICS. William Newman and Robert Sproull. McGraw-Hill Book Co., 1973.

DEUTERIUM IN THE UNIVERSE

NUCLEAR ASTROPHYSICS. William A. Fowler. The American Philosophical Society, 1967.

BIG-BANG NUCLEOSYNTHESIS REVISITED. Robert V. Wagoner in *The Astrophysical Journal,* Vol. 179, No. 2, pages 343–360; January 15, 1973.

ON THE ORIGIN OF THE LIGHT ELEMENTS. Hubert Reeves, Jean Audouze, William A. Fowler and David N. Schramm in *The Astrophysical Journal,* Vol. 179, No. 3, pages 909-930; February 1, 1973.

MATHEMATICAL GAMES

TRAVELERS IN TIME. Edited by Philip Van Doren Stern. Doubleday & Company, Inc., 1947.

SCIENCE FICTION ADVENTURES IN DIMENSION. Edited by Groff Conklin. Vanguard, 1953.

TIME AND FICTION IN DRAMA. *Man and Time.* J. B. Priestley. Doubleday & Company, Inc., 1964.

TIME AND THE SPACE-TRAVELLER. L. Marder. George Allen & Unwin, 1971.

THE AMATEUR SCIENTIST

METEOROLOGICAL SATELLITES. W. K. Widger, Jr. Holt, Rinehart and Winston, Inc., 1965.

AN INTRODUCTION TO ASTRODYNAMICS. Robert M. Baker, Jr., and Maud W. Makemson. Academic Press, 1967.

Our goal was to make getting there as much fun as being there.

When we designed our GMC MotorHome, we had something quite special in mind. We wanted a motorhome that would actually be fun to drive.

Which is why a GMC MotorHome looks the way it does. Aerodynamic. And low—lower than any other motorhome. So it can handle crosswinds. And corners.

That's also why we've provided it with front-wheel drive, a 455-cubic-inch V-8 with automatic transmission, and tandem rear wheels with air suspension. So it can handle highway speeds. And off-highway bumps.

Not that we've ignored living comfort. In fact, 15 floor plans, in four different color-coordinated decors, are available.

If you want specifics, send for our 28-page color catalog. Simply write: GMC MotorHome Headquarters, Drawer Y, Dept. 311, Lansing, Mich. 48909.

Better yet, see your local GMC MotorHome dealer.

The GMC MotorHome. It doesn't look like a box. It doesn't ride like a truck.

GMC

The MotorHome from General Motors.