Research Briefings 1986

Committee on Science, Engineering, and Public Policy
ISBN: 0-309-58176-1, 62 pages, 8.5 x 11, (1986)

Research Briefings 1986

Research Briefings 1986

For the Office of Science and Technology Policy, the National Science Foundation, and Selected Federal Departments and Agencies

Committee on Science, Engineering, and Public Policy
National Academy of Sciences
National Academy of Engineering
Institute of Medicine

NATIONAL ACADEMY PRESS
Washington, D.C. 1986

National Academy Press 2101 Constitution Avenue, NW Washington, DC 20418

NOTICE: The National Academy of Sciences was established in 1863 by Act of Congress as a private, nonprofit, self-governing membership corporation for the furtherance of science and technology for the general welfare. The terms of its charter require the National Academy of Sciences to advise the federal government upon request within its fields of competence. Under this corporate charter, the National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively.

The Committee on Science, Engineering, and Public Policy is a joint committee of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. It includes members of the councils of all three bodies.

This work was supported by the National Science Foundation under Grant LDA8501382.

Library of Congress Catalog Card Number 86-61858
International Standard Book Number 0-309-03689-5

Printed in the United States of America

Committee on Science, Engineering, and Public Policy

GILBERT S. OMENN, Dean, School of Public Health and Community Medicine, University of Washington, Seattle, Wash. *(Chairman)*

H. NORMAN ABRAMSON, Executive Vice-President, Southwest Research Institute, San Antonio, Tex.

FLOYD E. BLOOM, Director and Member, Division of Pre-Clinical Neuroscience and Endocrinology, Scripps Clinic and Research Foundation, La Jolla, Calif.

W. DALE COMPTON, Senior Fellow, National Academy of Engineering, Washington, D.C.

EMILIO Q. DADDARIO, Wilkes, Artis, Hedrick and Lane, Attorneys at Law, Washington, D.C.

GERALD P. DINNEEN, Vice-President, Science and Technology, Honeywell, Inc., Minneapolis, Minn.

RALPH E. GOMORY, Senior Vice-President and Director of Research, Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, N.Y.

ZVI GRILICHES, Professor of Economics, Harvard University, Cambridge, Mass.

ARTHUR KELMAN, Wisconsin Alumni Research Foundation Senior Research Professor of Plant Pathology and Bacteriology, Department of Plant Pathology, University of Wisconsin, Madison, Wis.

* PHILIP LEDER, John Emory Andrus Professor and Chairman, Department of Genetics, Harvard Medical School, Cambridge, Mass.

FRANCIS E. LOW, Institute Professor, Department of Physics, Massachusetts Institute of Technology, Cambridge, Mass.

EDWARD A. MASON, Vice-President, Research, Amoco Corporation, Amoco Research Center, Naperville, Ill.

* DANIEL NATHANS, Professor, Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, Md.

JOHN D. ROBERTS, Institute Professor of Chemistry, Gates and Crellin Laboratories of Chemistry, California Institute of Technology, Pasadena, Calif.

* Term expired June 30, 1986.

KENNETH J. RYAN, Kate Macy Ladd Professor of Obstetrics and Gynecology, Harvard Medical School, and Chairman, Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, Mass.

LEON T. SILVER, William M. Keck Foundation Professor of Geology, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, Calif.

HERBERT A. SIMON, Professor of Computer Science and Psychology, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pa.

* F. KARL WILLENBROCK, Cecil H. Green Professor of Engineering, School of Engineering and Applied Science, Southern Methodist University, Dallas, Tex.

Ex Officio

FRANK PRESS, President, National Academy of Sciences
ROBERT M. WHITE, President, National Academy of Engineering
SAMUEL O. THIER, President, Institute of Medicine.

COSEPUP Staff

ALLAN R. HOFFMAN, Executive Director
BARBARA A. CANDLAND, Administrative Coordinator
JOANNA MASTANTUONO, Senior Secretary

* Term expired June 30, 1986.

Preface

Research Briefings 1986 is the fifth volume of research briefing reports on selected areas of science and technology prepared by the Committee on Science, Engineering, and Public Policy (COSEPUP).* The briefings are prepared at the request of the President's Science Advisor, who also serves as Director of the White House Office of Science and Technology Policy (OSTP), and the Director of the National Science Foundation (NSF).

The four individual reports in this volume bring to 32 the number of such reports developed by COSEPUP. A list of the subjects covered in Research Briefings volumes each year since 1982 is appended to this preface. Together they constitute a rich set of assessments of recent advances and high-leverage research opportunities in a large number of critical fields of science and technology.

In addition to their collective value, individual briefing reports have had significant impact in their respective fields. For example, Dr. George Keyworth II (Presidential Science Advisor when the briefings were initiated) has noted that the research briefing on Computers in Design and Manufacturing "led almost directly and quickly to an important new program of Engineering Research Centers in the National Science Foundation." Other research briefings have influenced federal priorities and funding in mathematics, astronomy and astrophysics, plant sciences, and the solid earth sciences. Often, briefings have highlighted interdisciplinary connections, as in neuroscience, materials science, cognitive science and artificial intelligence, and biotechnology. Clearly, the research briefing activity has become an important new means of cooperation between the federal government and the National Academy complex.

As it has evolved through five annual rounds, the research briefing activity involves, first, the selection of topics by the OSTP and NSF directors in response to suggestions offered by COSEPUP. The briefings are then developed by panels of experts charged with assessing the status of the field and identifying those research areas within the field likely to return the highest scientific dividends as a result of near-term federal

* COSEPUP is a joint committee of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine, which together are known as the National Academy complex.

investment. COSEPUP then reviews the briefings before their presentation in oral and written form to officials in the White House, the National Science Foundation, other executive branch departments and agencies, and the Congress. The briefing reports, which are published and made widely available by the National Academy Press, are used in evaluations of the state of U.S. science and technology and in development and review of budget proposals.

Development of the briefings depends on the cooperative efforts of many: the scientists and engineers who serve in a volunteer capacity on the panels; those in the Science Advisor's office and the National Science Foundation who provide guidance and support throughout the process; and the National Academy complex staff who recommend potential topics and whose dedicated efforts facilitate the day-to-day activities of the briefing panels. To all, COSEPUP expresses deep appreciation and sincere thanks.

GILBERT S. OMENN, CHAIRMAN

COMMITTEE ON SCIENCE, ENGINEERING, AND PUBLIC POLICY

RESEARCH BRIEFING TOPICS*

1986

1. Science of Interfaces and Thin Films
2. Decision Making and Problem Solving
3. Protein Structure and Biological Function
4. Prevention and Treatment of Viral Diseases

1985

1. Remote Sensing of the Earth
2. Pain and Pain Management
3. Biotechnology in Agriculture
4. Weather Prediction Technologies
5. Ceramics and Ceramic Composites
6. Scientific Frontiers and the Superconducting Super Collider
7. Computer Vision and Pattern Recognition

1984

1. Computer Architecture
2. Information Technology in Precollege Education
3. Chemical and Process Engineering for Biotechnology
4. High-Performance Polymer Composites
5. Biology of Oncogenes
6. Interactions Between Blood and Blood Vessels (Including the Biology of Atherosclerosis)
7. Biology of Parasitism
8. Solar Terrestrial Plasma Physics
9. Selected Opportunities in Physics

1983

1. Selected Opportunities in Chemistry
2. Cognitive Science and Artificial Intelligence
3. Immunology
4. Solid Earth Sciences
5. Computers in Design and Manufacturing

1982

1. Mathematics
2. Atmospheric Sciences
3. Astronomy and Astrophysics
4. Agricultural Research
5. Neuroscience
6. Materials Science
7. Human Health Effects of Hazardous Chemical Exposures

* The reports listed here are published in *Research Briefings 1986, Research Briefings 1985,* etc., by the National Academy Press, Washington, D.C.

Contents

Report of the Research Briefing Panel on Science of Interfaces and Thin Films

Research Briefing Panel on Science of Interfaces and Thin Films

John A. Armstrong *(Cochairman),* IBM Corporation, Yorktown Heights, N.Y.
George M. Whitesides *(Cochairman),* Harvard University, Cambridge, Mass.
John H. Birely, Los Alamos National Laboratory, Los Alamos, N. Mex.
Peter R. Bridenbaugh, ALCOA Tech Center, Alcoa Center, Pa.
Norman Gjostein, Ford Motor Corporation, Dearborn, Mich.
Arthur C. Gossard, AT&T Bell Laboratories, Murray Hill, N.J.
Franz J. Himpsel, IBM Corporation, Yorktown Heights, N.Y.
N. J. Johnston, NASA Langley Research Center, Hampton, Va.
Robert W. Mann, Massachusetts Institute of Technology, Cambridge, Mass.
Thomas McGill, California Institute of Technology, Pasadena, Calif.
Calvin G. Quate, Stanford University, Stanford, Calif.
Dotsevi Y. Sogah, E. I. du Pont de Nemours & Company, Inc., Wilmington, Del.
Mark Wrighton, Massachusetts Institute of Technology, Cambridge, Mass.

Staff

William Spindel, *Project Director,* Commission on Physical Sciences, Mathematics, and Resources
Robert M. Simon, *Staff Officer,* Commission on Physical Sciences, Mathematics, and Resources
Alfred B. Bortz, *Consultant*
Sandra Nolte, *Senior Secretary*
Allan R. Hoffman, *Executive Director,* Committee on Science, Engineering, and Public Policy

Report of the Research Briefing Panel on Science of Interfaces and Thin Films

DEFINITIONS, PROPERTIES, AND BEHAVIORS

A *thin film* is matter of microscopic thickness—typically, only a few atoms to a few thousand atoms. Its extent in its other two dimensions is macroscopic. An *interface* is the junction of two different substances or two phases of the same substance. The properties of these quasi-two-dimensional entities are often remarkably different from the properties of bulk matter of the same composition.

Thin films and interfaces are associated concepts. Thin films (for example, the iridescent film that forms when oil floats on water) are familiar. Interfaces are less familiar, although ubiquitous; they occur wherever different homogeneous phases meet. For example, the oil film on water has two interfaces: one with water and one with air. As the thickness of a film decreases, its properties are increasingly determined by its interfaces. At the limit of thinness, thin films and interfaces merge. An interface can be thought of as a film so thin it has no homogeneous interior; a thin film is a system whose interior is strongly influenced by the close proximity of its interfaces.

The current interest in thin films and interfaces reflects both timely opportunities in basic science and importance and pervasiveness in technology.

- A broadly important problem in science is that of the basic relations between the macroscopic properties of matter (e.g., wettability, electrical and thermal conductivity, hardness, reflectivity) and its atomic-level structure. Because the detailed structures of interfaces and thin films can be manipulated with greater control than can those of bulk solids or liquids, they provide particularly attractive systems for studying these basic relations.
- A number of interesting phenomena—especially manifestations of quantum behavior—only appear in small systems. Thin films and interfaces can be of a size that displays these phenomena. Thus, for example, the rate of transport of electrons across thin films by quantum tunneling may be very high when the thickness of the films is comparable to the extension of electronic wave functions (10 to 100 angstroms).
- The structure, properties, and reactivity of matter at an interface can be very different from those of matter in bulk because of

the close proximity of the interfacial matter to matter of different composition, or of the interfacial matter to vacuum. Thus, gold atoms at a gold-silicon interface do not behave like gold atoms in bulk because many of their nearest neighbors are silicon atoms; gold atoms at a gold-vacuum interface behave differently from those in bulk because they lack the complement of near neighbors present in a solid.

- The connections between science and technology are particularly close and useful in matters concerning thin films and interfaces. Technologically important physical properties—strength, corrosion resistance, biocompatibility—are often determined by the characteristics of thin films and interfaces.

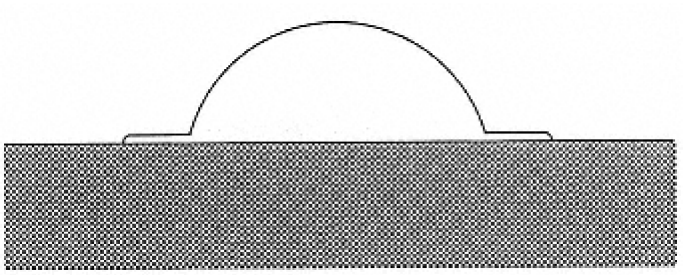

Figure 1
Schematic cross section of a drop of water spreading on a solid support. The existence of a "precursor film," a thin lip of liquid (not drawn to scale) extending beyond the drop, is due to long-range interactions between the solid and the liquid. Source: Based on P. G. deGennes, Reviews of Modern Physics, Vol. 57, No. 3 (1985):827–863.

An interface can be the exposed atoms on the exterior of a metal single crystal in vacuum, the junction between a silicon substrate and a silicon dioxide overlayer, the boundary between phase-separated block copolymers, the junction between a fiber and polymer matrix in a composite aircraft part, or the region of contact between blood and an implanted prosthesis. A thin film can be the membrane enclosing a living cell; the thin layers of conductors, semiconductors, and insulators that constitute a microelectronic device; the layered media of a magnetic information storage disk; the layers used for lubrication, adhesion, and passivation; or the coatings of surfactants that are used to stabilize suspensions. The properties of these thin films depend strongly on their constituent interfaces.

The examples that follow illustrate several interesting and important characteristics of matter in two dimensions.

- *The presence of a solid in contact with liquid water has a profound influence on the character of the water.* Figure 1 is a diagram of a small drop of water spreading on a solid. The curvature of the major part of the surface of the drop is determined by a balance of energies at the liquid-vapor, liquid-solid, and solid-vapor interfaces. A feature of great current interest is the so-called "precursor film," a lip of liquid a few hundred angstroms thick extending for microns beyond the edge of the drop. Current explanations of this precursor film attribute its existence, at least in part, to long-range interactions between the liquid and the solid surface. The strength of these interactions is sufficient to pull a thin film at the edge of the drop flat against the surface. Electrochemical evidence supports a model for water next to an interface that is qualitatively different from bulk water: interfacial water may have a dielectric constant as low as 30 (the dielectric constant of bulk water is 78). Understanding the behavior of liquid-solid interfaces is critical to understanding wetting, and thus to such technologies as adhesion and corrosion protection.
- *Matter in thin films may exhibit phase behavior that is quite different from the phase behavior the same matter exhibits in bulk.* A system composed of krypton that is adsorbed on

graphite at a pressure and temperature sufficient to give a coverage of one to two monolayers shows remarkably complex phase behavior. The krypton in the first monolayer is a fluid at 130K, the high end of the temperature range. As the temperature is lowered, the krypton first freezes into a two-dimensional solid, then melts into a new fluid, and finally freezes again into a new solid. The structures of the solid phases exhibited by this system have been characterized (see Figure 2). In the higher-temperature solid phase, the krypton atoms position themselves in a low-density crystal in register with the underlying graphite lattice. This structure is dominated by krypton-carbon interactions; and in it, the krypton atoms are spaced slightly beyond hard-sphere contact. In the lower-temperature solid, krypton condenses to a higher density in the first monolayer and crystallizes in a structure dominated by krypton-krypton interactions; in this structure, the registration with the graphite lattice is destroyed. This system illustrates the delicate balance between forces among the krypton atoms in an adsorbed monolayer and between these atoms and the graphite substrate on which they are adsorbed. Studies such as this of krypton on graphite establish the ability of current instrumental techniques to characterize the structures of monolayers; they also provide a starting point for studying the thermodynamics of thin films. Although inert gas systems are of limited practical interest, they are the simplest systems to analyze theoretically, and they provide conceptual models for systems dominated by weak atom-atom interactions.

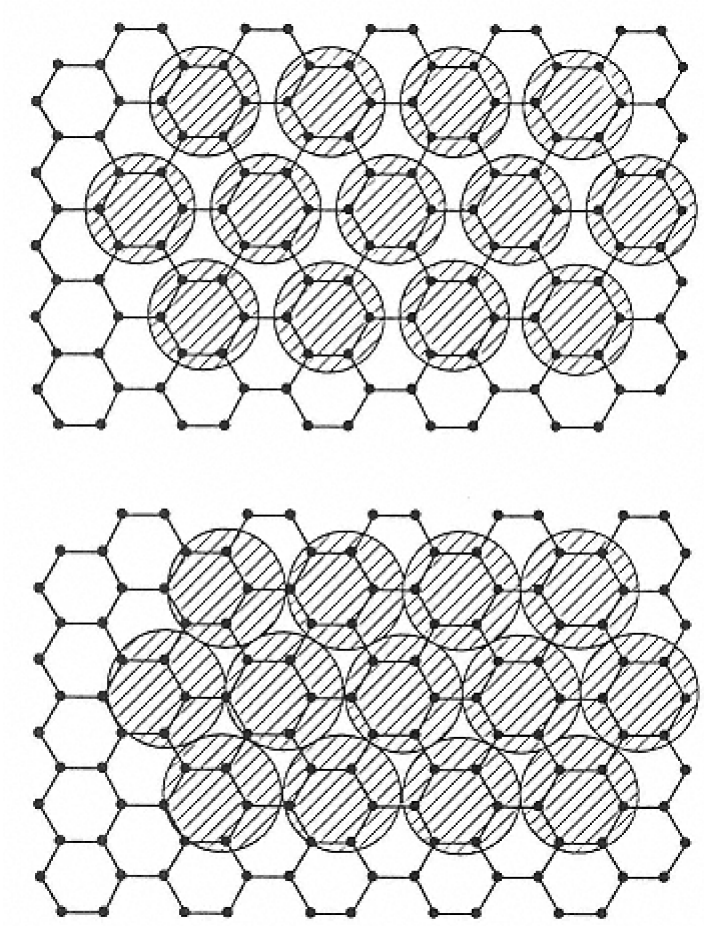

Figure 2
Krypton adsorbed on graphite. A low density crystalline phase (upper diagram) is dominated by krypton-graphite interactions: the krypton atoms are placed slightly beyond hardsphere contact. The high-density crystalline phase (lower diagram) brings the krypton atoms into contact but destroys the registration with the underlying graphite lattice.
Source: Based on R. J. Birgeneau and P. M. Horne, Science 232 (1986):329.

- *Bulk and interfacial electrical and magnetic properties of materials may be strikingly different.* Bulk crystalline silicon is a semiconductor; the interface of a silicon crystal cut to expose a particular crystal plane, the Si(111) face, exhibits metallic behavior. This phenomenon is not yet understood theoretically, but a critical first step—establishing the structure of the Si(111) interface—has been taken recently through the use of scanning tunneling microscopy. The observed structure shows significant changes in atomic positions relative to bulk; some bond angles are different.
- *Far from being passive containers for the contents of the cell, the membranes covering cells are highly organized, dynamic, structurally complex biological systems that regulate communication between matter lying inside and outside of the cells.* One important constituent of cell membranes is a class of molecules, the phospholipids, that spontaneously forms bilayer films in a number of geometries. Many of the important physical properties of cell membranes, such as two-dimensional diffusion and differentiation between the "inside" and "outside" of biological entities shaped like a tube or sphere, can be studied using these spontaneously formed structures.
- *The characteristic chemical reactivities of metal atoms at the exposed interface of bulk metal and of small metal clusters provide the basis for heterogeneous catalysis.* Heterogeneous catalysis (that is, catalysis using solid catalysts) is an important technology for the production of fuels and chemicals. The metal atoms used in a heterogeneous catalyst are chosen for their high reactivity. As a result of their location in a solid interface, they are simultaneously accessible to reactants in a contacting vapor or liquid and isolated from reaction with one another.
- *Microelectronic devices are assemblies of thin films; many of the properties of these devices derive from the special properties of electrons in the films and the transport of electrons in and across the interfaces joining them.* Figure 3 is a schematic cross section through one part of such a device: a multilayer system connecting a transistor to make ohmic contact to a solder pad. The conductors in this system are 0.1 to 5 micrometers (1,000 to 50,000 angstroms) in cross section; in operation, they can carry current densities of up to 10^5 amps/cm^2. The force of this "electron wind" is sufficient to cause electromigration, or migration of atoms in the conductors from their normal lattice sites. Remarkably, some of these atoms migrate with the wind, and some migrate against it. The basis of this phenomenon is incompletely understood, but its origin clearly lies in the small size of the conductors; its control is important in reducing the size of integrated circuit chips.
- *Insertion of a monolayer, 20 angstroms thick, of a simple organic substance—hexadecylamine, $CH_3(CH_2)_{15}NH_2$*—between two steel surfaces in sliding contact reduces the friction between them by a factor of 10. This reduction in friction reflects spontaneous formation of an ordered film of the organic species: the amine (NH_2) groups bond to the metal, and the hydrocarbon chains orient roughly perpendicular to it. The film is thus a thin hydrocarbon liquid or liquid crystal bonded to the metal, a material that resists the transitory adhesion between the metal pieces that contributes strongly to friction. Understanding the details of the relations between the structure of the adsorbing organic molecules, the solid phase on which they adsorb, and the structure and properties of these types of spontaneously self-organizing monolayer films promises to stimulate the design of thin-film systems to control friction, wear, corrosion, and adhesion.

In short, matter present in thin films or at interfaces can exhibit unique properties. Understanding and controlling these properties have been difficult because of the small quantities of material present in most interfaces and thin films relative to the bulk, and because many of the interfaces of greatest interest are "buried" inside solids or under

liquids. New instrumental techniques combined with theory and computer modeling, however, enable us to define the structures of many interfaces. With sound structural information and new methods of preparation, it is now possible to explore the rich phenomenology of interfaces and thin films.

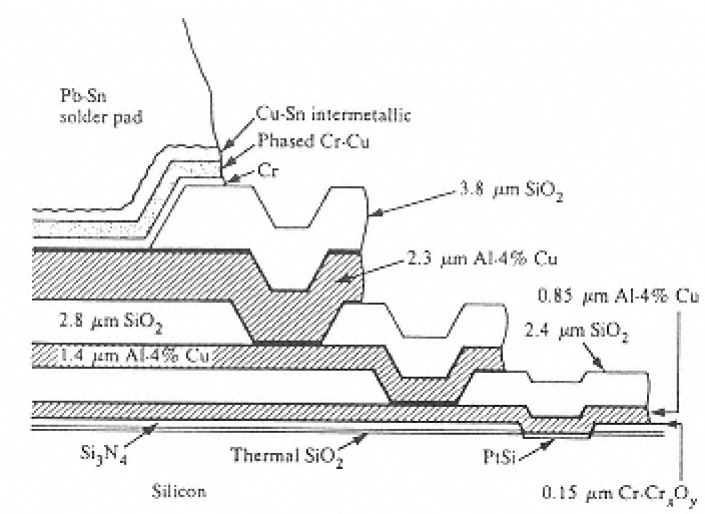

Figure 3
Sectional drawing of multilevel interconnections for advanced bipolar devices. Source: L. J. Fried et al., IBM Journal of Research and Development 26 (3 May 1982).

RESEARCH ISSUES

Characterization of Interfaces and Thin Films

The bonding characteristics and electronic structure of most interfaces (even the "simple" solid-vacuum interfaces of crystalline elements) are still poorly understood, and no technique for establishing these structures is universally applicable. (The structure of an interface cannot necessarily be extrapolated from that of the underlying bulk solid.)

Much of the information about the structures of interfaces has come from forms of spectroscopy that are limited to solid-vacuum interfaces, but emerging techniques can now characterize solid-gas and solid-liquid interfaces. The most exciting of these techniques is *scanning tunneling microscopy,* or STM (Figure 4). STM measures the very small current that flows when a potential is applied between a conducting interface and a probe tip (only a few atoms in size) scanned across the interface at a distance of angstroms. The current is caused by quantum tunneling of electrons between individual atoms on the interface and on the probe, and it is extremely sensitive to the distance between the atoms. This remarkable device makes it possible to observe individual atoms on irregular interfaces. STM is being used to study interfaces in contact with insulating liquids; it is applicable to noncrystalline solids; and it can be used to examine dynamic processes occurring at interfaces. It should be particularly useful in examining the structures of individual defects on interfaces.

A second instrument relying on the ability to position interfaces with angstrom-scale control is the *interface force balance.* This device holds two flat solids (for example, sheets of mica) at accurately known separations of from 3 to 500 angstroms, and measures the attraction or repulsion between them. The measurements can be carried out with the solids separated by vacuum, gas, or liquid, or with the solids carrying monolayers of other materials bonded or adsorbed to their interfaces. The measurements make possible direct analysis of the forces responsible for interactions between interfaces, modification of these forces by intervening con

densed phases, and, by inference, interaction of the interfaces with the condensed phases.

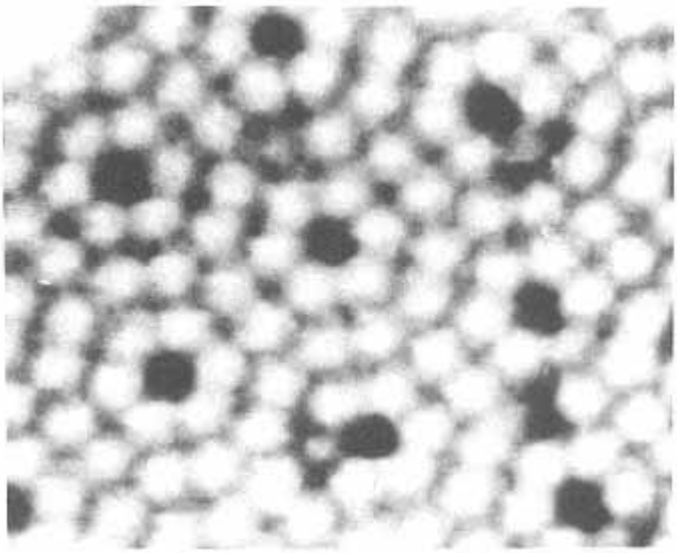

Figure 4
Scanning tunneling microscopy spectroscopy—a topograph of an Si(111) 7 × 7 surface. By changing the voltage, one is able to measure electronic states within an area of atomic dimensions. Defects (some of which show up as extra dark spots in the topograph) can also be probed by this method.
Source: J. E. Demuth et al., IBM Corporation.

A selection of the wide range of other instrumental techniques now being used to characterize interfaces is presented in Table 1. The sheer number of available techniques presents interface science with both an opportunity and a problem. The techniques provide many useful and complementary types of information; but because no single technique uniquely characterizes any system, it is necessary for an effective laboratory to have access to several instruments. Their expense and complexity in turn raise a substantial problem in management for small research groups.

Current objectives of research in structural aspects of interface science are the development of techniques for characterizing buried interfaces (for example, grain boundaries in metals and ceramics, the fiber—matrix interface, and defects in composites); and for examining electrically insulating interfaces such as those on organic polymers.

Preparative Techniques

One objective of current research is the development of techniques for producing highly perfect, smooth, single-crystal films. A number of techniques are used for preparation of electronic materials, ranging from simple vacuum evaporation to *molecular beam epitaxy* (MBE) and *metal-organic vapor phase epitaxy* (MOVPE). *(Epitaxy* is the growth of a crystalline film of one material on a crystal face of a second in such a way that the crystalline orientation of the deposited material follows that of the substrate.) Current scientific understanding of the processes underlying all these techniques—processes that include the movement of species on the interfaces, mechanisms of annealing and relief of stress, incorporation of impurities, nucleation and growth of defects—is limited; an improvement in that understanding would be immensely valuable in technology.

Epitaxial growth techniques seem certain to be particularly useful. They can be used to make very small structures such as quantum wells (structures whose composition is tailored at angstrom scales to control electronic energy levels); films with extraordinarily high electron and hole mobilities; and transistors, lasers, and magnetic materials that are capable of record-setting performances.

The most interesting strategies for preparation of thin films of organic constituents

are based on the spontaneous self-assembly of low molecular weight molecules at interfaces. For example, so-called Langmuir-Blodgett monolayers are formed by spreading an organic substance such as stearic acid ($CH_3[CH_2]_{16}CO_2H$) at the interface between air and water. The hydrophilic CO_2H groups are attracted to the water, while the hydrophobic hydrocarbon chains are excluded from it. When the surface film is compressed, the organic molecules pack as a two-dimensional crystal or liquid crystal that can be transferred intact to a solid support. Similar films can be formed in many cases by simply allowing the organic substance to adsorb from solution onto a support: the desired ordering occurs spontaneously. These techniques make it possible to prepare macroscopic, two-dimensional, organic monolayer films while maintaining a high degree of control over the nature of the exposed surface functional groups, the order of the films, and (to a more limited extent) their physical and mechanical properties. These systems are exceptionally attractive as substrates for the study of relationships between interface structure and such properties as wettability, biocompatibility, electrical resistivity, and nonlinear optical response.

Properties of Matter in Two Dimensions

The quasi-two-dimensional geometry of interfaces and thin films, and the high gradients in properties across them, can result in unique properties for matter in these systems. Of the wide range of topics that might be used to illustrate the characteristic properties of interfaces and thin films, device physics offers a particularly clear demonstration of the interplay between science and technology. The use of thin films is important in the construction of devices for two reasons: (1) small size permits a high density of devices, thus minimizing power consumption and maximizing speed; and (2) small size is required for many devices that exploit quantum effects.

TABLE 1 Selected Spectroscopic Techniques Applicable to Interfaces

Technique (Acronym)	Application
Scanning tunneling microscopy (STM)	Individual atomic positions on surfaces
Neutron scattering Low-angle x-ray scattering	Structures of crystalline surfaces; surface morphology
Surface-enhanced Raman spectroscopy (SERS)	Vibrational spectroscopy of adsorbates on small metal particles
X-ray photoelectron spectroscopy (XPS); Auger spectroscopy	Electronic structure of atoms in the top 10 angstroms of an interface
High-resolution electron microscopy	A wide variety of information concerning structure, composition, and morphology; single atom imaging in special cases
Rutherford backscattering (RBS)	Atomic composition as a function of depth with resolution of hundreds of angstroms
Secondary ion mass spectroscopy (SIMS)	Molecular-sized fragments of interfaces
Reflectance infrared spectroscopy	Vibrational and structural analysis of organic thin films
Acoustic microscopy	Interface structures at the 10,000 Å level
Nuclear magnetic resonance spectroscopy (NMR)	Interface structures at the 10,000 Å level
Electron paramagnetic resonance spectroscopy (EPR)	Paramagnetic centers in interfaces

A two-dimensional electron gas exists at the interface between the silicon channel and gate insulator in metal oxide semiconductor (MOS) devices; similar electron gases are

important in many other types of devices. The transport properties of these two-dimensional electron gases exhibit a number of new phenomena: for example, negative resistance, in which current decreases with increases in applied voltage because of electron tunneling into energy sub-bands with lower momentum; and ballistic transport, in which electrons move without phonon scattering in structures with very small dimensions. *Heterojunctions* (structures involving an interface between two different semiconductors) are the basis for a number of developments in device physics. New light-emitting structures involving multiple heterojunctions (so-called *quantum-well light emitters)* provide light at wavelengths previously unattainable. *Superlattices* formed by producing a series of periodically spaced heterojunctions are materials with new non-linear optical and magnetic properties.

Interface Reactivity

Atoms and molecules present at an interface can experience a highly anisotropic environment with characteristics that are different from surrounding bulk phase(s). The chemical reactivity of a species present at an interface may, in consequence, be different from the reactivity of the same species present in an isotropic phase.

As one example, aggregates of platinum atoms supported on alumina react with hydrocarbons in ways that depend strongly on the size and shape of the metal aggregate, on the acidity of the underlying alumina support, and on the nature of the interaction between the aggregate and the support. Platinum atoms are intrinsically highly reactive toward hydrocarbons. But platinum atoms in bulk platinum are not accessible; hence, they are not active catalytically. Platinum atoms in solution tend to react indiscriminately with one another, with species used to increase their solubility, and with the intended reactants. Small aggregates of supported platinum thus provide systems in which a high proportion of the platinum is stably isolated and exposed at an interface, available for reaction. In these systems, the reactivity of the platinum can also be tailored to favor useful reactions by changing the underlying support. The reactivity toward hydrocarbons of platinum supported on alumina forms the basis for one of the critical steps in petroleum refining.

A second example is the unexpected acidities of organic functional groups present at the interface between low dielectric polymers, such as polyethylene, and water. The apparent acidity of a carboxylic acid (CO_2H) group at such an interface—that is, the concentration of protons in solution at which these groups are half-ionized to carboxylate ions (CO_2^-)—is shifted by 10^6 from that characterizing the same groups in homogeneous aqueous solution. The shift in apparent acidity reflects three factors: (1) the low local dielectric constant at the polymer-water interface; (2) the anomalously low polarity of the water present at this interface; and (3) electrostatic interactions at the surface. The unexpected reactivity of functional groups present at polymer—fluid interfaces is clearly relevant to wetting of and adhesion to polymers, and to other processes involving the reaction and solvation of functional groups present in polymer interfaces. It is also relevant to the characteristics of functional groups present at many other interfaces, especially those between water and suspensions, micelles, and proteins.

Biocompatible Surfaces and Interfaces

A biomaterial is any substance or device whose function depends on contact with a biological medium. Thin films and interfaces play an essential role in the design and function of the numerous implants and devices that are now being used clinically. Their surfaces induce deposition of proteins, platelets, and other cellular elements, and

often induce platelet aggregation and blood clot (thrombus) formation.

Although functional aspects of the performance of artificial materials in the human body can be predicted with some reliability, forecasting their biological performance is difficult. Fundamental information on the correlation between the in vivo and in vitro responses is limited. An understanding of the dynamic biological changes occurring at the material—tissue interface is necessary to predict biological performance. For practical application, factors influencing interactions of blood with materials and biological surfaces are particularly important. We must learn how these interactions are affected by surface morphology or by specific chemical groups on the surface, and what influence is exerted by the mechanical properties of the interface.

APPLICATIONS

Microelectronics in Communications and Computers

Microelectronic technologies are founded largely on structures composed of thin films and interfaces. The silicon-based semiconductor industry depends critically on the still incompletely understood interface between silicon and insulators such as silicon dioxide and silicon nitride. Lasers, detectors, and new high-speed devices use heterojunctions in which the interface is the key active component. The conduction of electricity in microfabricated devices requires reproducible interfaces between metallic conductors and the active semiconducting layer, the so-called ohmic contact. Such contacts become increasingly difficult to establish as dimensions decrease. Mass storage technologies (magnetic disks, optical and magneto-optical storage devices, magnetic bubble memories) depend on films that are only hundreds of angstroms in thickness but that are uniform over many square centimeters.

Packaging and interconnect technologies are as important as chip technologies to future developments in high-performance computing systems. Processing speeds are often limited by the time required to propagate information from one chip or subsystem to another, and the fabrication and assembly of very small systems is crucial in building high-speed computers. Of necessity, the components in these systems are densely packed, and they may generate large quantities of heat whose removal is controlled by interfacial thermal conductivity.

Long-term objectives include the development of systems that allow direct connection of electronic devices to biologically based sensors or to nerves. Both require the solution of substantial problems in interface science. Another challenging interfacial problem is the development of practical techniques that interconvert electrons (the currency used by digital devices) and neurotransmitters (the currency used by nerves) to permit the nondestructive stimulation or sensing of nerve impulses.

Control of Corrosion, Friction, and Wear

Corrosion is the result of electrochemical processes occurring at interfaces between metal and water and air. Control of corrosion is usually achieved by the application of thin protective films to the metal interface. Although techniques for corrosion control are highly developed empirically, too often they are effective only for short periods, and their fundamental basis is often obscure. Studies of the formation, thermodynamic and kinetic stability, and barrier properties of thin films on metals are now possible at levels of detail that will be increasingly useful in developing new strategies for control of corrosion.

The practice of lubrication has developed adequate engineering models for elastohydrodynamic lubrication (EHL), provided the lubricant is present as a thick film and be

haves as a Newtonian fluid. As the thickness of the lubricant film approaches molecular dimensions, as in asperity contact, pressures become high, approaching 1 gigapascal; shear forces become very high; and lubricants solidify, crystallize, and degrade. Many lubricated surfaces are covered with softer adherent films derived at least in part from the lubricant; these films mitigate asperity contact when forces exceed the EHL limit. Understanding thin lubricant and adherent films, especially under the extreme conditions of asperity contact, is now conceivable, although still difficult.

Controlling wear at the head-medium interface in a high-density magnetic storage device provides an example of a current problem in this technology. The thin-film magnetic recording medium on a disk is part of a multilayer structure. It is protected by another thin film—a wear-resistant overcoat—that may, in the future, be a diamond-like carbon. The magnetic medium may be bonded to its substrate by yet another thin film, and there may be a magnetic underlayer. In Winchester hard disk technology, a microfabricated head literally flies over the disk, supported by an air bearing whose width is comparable to the average distance between collisions of individual molecules in ambient air (~900 angstroms). Efforts to increase the density of data storage require closer head/disk spacings; they also raise formidable problems in lubrication, "stiction" (adhesive effects of the lubricant during static head/disk contact), wear, adhesion, and delamination.

Structural Materials

Interfaces play a central role in determining the mechanical properties of a range of structural materials, among which are fiber-reinforced composites, metals, and ceramics. In broad terms, the failure of structural elements almost always involves inelastic deformation and fracture. The atomistics—the details of kinetics and thermodynamics—of the processes involved in initiation and propagation of cracks in a material under strain, and the processes that protect against fracture (dissipation of energy through the creation of dislocations, formation of cracks and voids, and other mechanisms) are various and incompletely understood. But they all have as a common result the formation or alteration of interfaces.

Fiber-reinforced composites consist of an ordered dispersion of fibers in a polymer matrix. The fibers provide stiffness and strength; the matrix distributes the load among the fibers and protects them from damage. The large area of interface connecting fiber and matrix makes a critical contribution to the mechanical properties of the material. Failure often involves these interfaces (Figure 5), but the relationship between their structure and mechanical performance is poorly understood. How tightly should the fiber and the matrix adhere to achieve optimal performance? What mechanical properties in the interface best dissipate energy during incipient failure? How should the optimum interface be prepared? What is the proper surface chemistry for the fiber, and how should the fiber be brought into contact with the matrix? Answers to these questions would provide the basis for rational optimization of composite systems and would be of substantial value to users of composites, especially in the automotive and aircraft industries.

Energy Production

Chemically active interfaces are required as heterogeneous catalysts in the refining of petroleum. *Corrosion-resistant interfaces* are needed in the heat transfer systems of power plants.

Fuel cells and certain kinds of batteries can be substantially improved by *new catalytic electrodes* that will allow more efficient use of atmospheric oxygen. The reduction of oxygen by electrons to water is slow at

conventional electrodes, and even the best platinum catalysts have rates that give output voltages that are only one-half of what is theoretically possible. The inefficiency of these catalysts is not intrinsic to all systems that reduce oxygen; in biological systems, through molecular mechanisms that are still incompletely understood, the reduction of oxygen to water takes place rapidly. Recent experiments with synthetic catalysts have demonstrated the reduction of oxygen to water without the use of platinum, but substantial improvements in the lifetime and cost of electrode-confined catalysts must be achieved before the systems are economical.

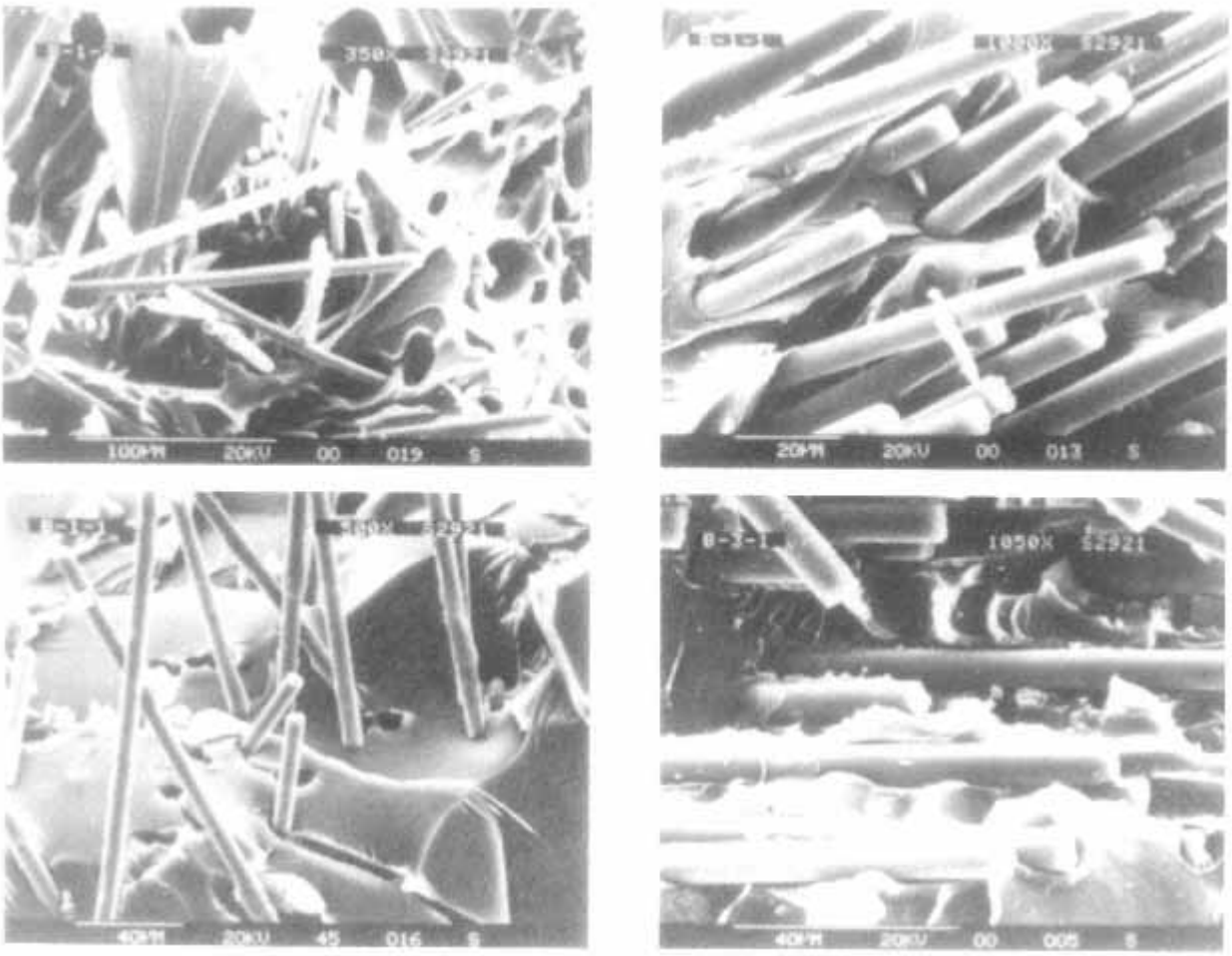

Figure 5
Fractured surface of a polymeric composite: unsized chopped carbon fibers in a polycarbonate matrix.
Source: NASA Jet Propulsion Laboratory.

The production of petroleum illustrates another important set of interfacial problems. One of the approaches to increasing the production of crude oil from partially depleted fields is to "launder" the oil-bearing rock using detergents similar to those used in cleaning oil-stained clothes. Water containing the detergents is pumped into the reservoir; the detergent separates the petroleum from the rock and permits it to pass through fine pores in the rock.

Currently, the economic feasibility of detergent-based methods for oil recovery remains unclear for most oil fields; it has proved difficult to find detergent mixtures that are inexpensive, effective, and stable in the res

ervoir. Designing successful detergents will require a detailed knowledge of the relations between their structures and the ways in which they modify the properties of oil-water interfaces.

National Security

Many of the applications of interface and thin-film technology in military and civilian systems are similar, with the important difference that military systems must function in extremely demanding environments with very high standards of performance. A fiber-reinforced composite wing for a military aircraft is subjected to greater stress than a similar wing on a civilian aircraft; therefore, optimization of the fiber—matrix interface is correspondingly more important. Because an engine for a tank operates much closer to its limits of failure than a comparable engine for a civilian truck, the design of critical components such as bearings requires better control of interfaces to minimize friction and wear.

There are also, however, certain areas of technology in which military requirements are unique. Interfaces are critical wherever high electromagnetic fields interact with matter. Mirrors for high-powered *lasers* must be immune to optical damage, and interfacial phenomena play a dominant role in determining their damage thresholds. Multilayer structures composed of alternating thin films of refractory metals and dielectrics are important for *x-ray optics.* High-powered accelerators operate with very high surface fields in their resonant cavities; the composition and morphology of the cavity interfaces contribute both to the sharpness of the resonant frequency and to the rate of surface heating during use. Certain military electronics devices must resist *damage by radiation;* high speeds and low power consumption are also important. Promising technologies to meet these requirements are based on thin-film heterostructure devices. In addition, new thin-film coatings are needed that prevent the reaction of lithium hydride and uranium with water vapor over intervals of decades. The *stabilities of nuclear devices,* an important element in the design of test ban treaties, are greatly influenced by interfacial reactivities.

MANAGEMENT ISSUES

Increased U.S. investment in the science of thin films and interfaces will produce a large return in improved technology. What are the goals of an appropriate investment strategy? What are the management issues raised by its execution?

1. Investment should build strong, two-way interactions between basic science and advanced technology. Inefficient two-way communication between these two spheres is a major hindrance in the cycle of product development; because technological application of thin films and interfaces is only a small step beyond basic research, research results will be invaluable to technologists. Unexplained and uncontrolled phenomena in technology, in turn, will stimulate basic research.
2. Investment should encourage interdisciplinary collaboration and join and exploit the strengths of academic, industrial, and government research and development institutions.
3. An effective strategy should provide selected groups and/or institutions with a sufficiently complete subset of the sophisticated analytical and preparative tools required to conduct effective research in thin films and interface science. One possible institutional structure would be research groups of about four faculty members that would focus on a coherent theme (for example, microelectronic interfaces, electroactive surfaces, or biocompatible materials). Each such group would manage a reasonable subset of preparative and analytical tools (see Table 1) that while possibly incomplete should still be sufficient to carry out a major

portion of the group's preparative and analytical work, supplemented by the use of equipment in other locations.

4. The scale of modern surface and interface research falls in between “small” and “big” science; this intermediate size complicates issues in management. For example, although individual researchers are not forced (as is the case in experimental high-energy physics) into large collaborative projects with explicit management structures, the single principal investigator, with his specialized technique or knowledge, seldom has adequate resources to solve complex experimental problems.
5. Personnel Requirements. Too few trained students are being produced in some important areas of interface research. For example, although there are many well-trained students in compound semiconductor interface science, there are comparatively few in such areas as polymer-metal, polymer-carbon, and polymer-ceramic interfaces; silicon epitaxial growth; colloid science; and biocompatible surfaces.

In summary, interface science is exceptional in its close, two-way interaction with technology. Only within the past few years have the tools and techniques become available for a concerted attack on the scientific problems of buried interfaces. Increased investment is therefore both scientifically timely and certain to pay significant dividends for technology.

Report of the Research Briefing Panel on Decision Making and Problem Solving

Research Briefing Panel on Decision Making and Problem Solving

Herbert A. Simon *(Chairman),* Carnegie-Mellon University, Pittsburgh, Pa.
George B. Dantzig, Stanford University, Stanford, Calif.
Robin Hogarth, University of Chicago, Chicago, Ill.
Charles R. Plott, California Institute of Technology, Pasadena, Calif.
Howard Raiffa, Harvard Business School, Boston, Mass.
Thomas C. Schelling, Harvard University, Cambridge, Mass.
Kenneth A. Shepsle, Washington University, St. Louis, Mo.
Richard Thaler, Cornell University, Ithaca, N.Y.
Amos Tversky, Stanford University, Stanford, Calif.
Sidney Winter, Yale University, New Haven, Conn.

Staff

David A. Goslin, *Executive Director,* Commission on Behavioral and Social Sciences and Education
Karan Ford, *Administrative Secretary*
Allan R. Hoffman, *Executive Director,* Committee on Science, Engineering, and Public Policy

Report of the Research Briefing Panel on Decision Making and Problem Solving

INTRODUCTION

The work of managers, of scientists, of engineers, of lawyers—the work that steers the course of society and its economic and governmental organizations—is largely work of making decisions and solving problems. It is work of choosing issues that require attention, setting goals, finding or designing suitable courses of action, and evaluating and choosing among alternative actions. The first three of these activities—fixing agendas, setting goals, and designing actions—are usually called *problem solving;* the last, evaluating and choosing, is usually called *decision making*. Nothing is more important for the well-being of society than that this work be performed effectively, that we address successfully the many problems requiring attention at the national level (the budget and trade deficits, AIDS, national security, the mitigation of earthquake damage), at the level of business organizations (product improvement, efficiency of production, choice of investments), and at the level of our individual lives (choosing a career or a school, buying a house).

The abilities and skills that determine the quality of our decisions and problem solutions are stored not only in more than 200 million human heads, but also in tools and machines, and especially today in those machines we call computers. This fund of brains and its attendant machines form the basis of our American ingenuity, an ingenuity that has permitted U.S. society to reach remarkable levels of economic productivity.

There are no more promising or important targets for basic scientific research than understanding how human minds, with and without the help of computers, solve problems and make decisions effectively, and improving our problem-solving and decision-making capabilities. In psychology, economics, mathematical statistics, operations research, political science, artificial intelligence, and cognitive science, major research gains have been made during the past half century in understanding problem solving and decision making. The progress already achieved holds forth the promise of exciting new advances that will contribute substantially to our nation's capacity for dealing intelligently with the range of issues, large and small, that confront us.

Much of our existing knowledge about decision making and problem solving, derived from this research, has already been

put to use in a wide variety of applications, including procedures used to assess drug safety, inventory control methods for industry, the new expert systems that embody artificial intelligence techniques, procedures for modeling energy and environmental systems, and analyses of the stabilizing or destabilizing effects of alternative defense strategies. (Application of the new inventory control techniques, for example, has enabled American corporations to reduce their inventories by hundreds of millions of dollars since World War II without increasing the incidence of stockouts.) Some of the knowledge gained through the research describes the ways in which people actually go about making decisions and solving problems; some of it prescribes better methods, offering advice for the improvement of the process.

Central to the body of prescriptive knowledge about decision making has been the theory of subjective expected utility (SEU), a sophisticated mathematical model of choice that lies at the foundation of most contemporary economics, theoretical statistics, and operations research. SEU theory defines the conditions of perfect utility-maximizing rationality in a world of certainty or in a world in which the probability distributions of all relevant variables can be provided by the decision makers. (In spirit, it might be compared with a theory of ideal gases or of frictionless bodies sliding down inclined planes in a vacuum.) SEU theory deals only with decision making; it has nothing to say about how to frame problems, set goals, or develop new alternatives.

Prescriptive theories of choice such as SEU are complemented by empirical research that shows how people actually make decisions (purchasing insurance, voting for political candidates, or investing in securities), and research on the processes people use to solve problems (designing switchgear or finding chemical reaction pathways). This research demonstrates that people solve problems by selective, heuristic search through large problem spaces and large data bases, using means—ends analysis as a principal technique for guiding the search. The expert systems that are now being produced by research on artificial intelligence and applied to such tasks as interpreting oil-well drilling logs or making medical diagnoses are outgrowths of these research findings on human problem solving.

What chiefly distinguishes the empirical research on decision making and problem solving from the prescriptive approaches derived from SEU theory is the attention that the former gives to the limits on human rationality. These limits are imposed by the complexity of the world in which we live, the incompleteness and inadequacy of human knowledge, the inconsistencies of individual preference and belief, the conflicts of value among people and groups of people, and the inadequacy of the computations we can carry out, even with the aid of the most powerful computers. The real world of human decisions is not a world of ideal gases, frictionless planes, or vacuums. To bring it within the scope of human thinking powers, we must simplify our problem formulations drastically, even leaving out much or most of what is potentially relevant.

The descriptive theory of problem solving and decision making is centrally concerned with how people cut problems down to size: how they apply approximate, heuristic techniques to handle complexity that cannot be handled exactly. Out of this descriptive theory is emerging an augmented and amended prescriptive theory, one that takes account of the gaps and elements of unrealism in SEU theory by encompassing problem solving as well as choice and demanding only the kinds of knowledge, consistency, and computational power that are attainable in the real world.

The growing realization that coping with complexity is central to human decision making strongly influences the directions of research in this domain. Operations research and artificial intelligence are forging

powerful new computational tools; at the same time, a new body of mathematical theory is evolving around the topic of computational complexity. Economics, which has traditionally derived both its descriptive and prescriptive approaches from SEU theory, is now paying a great deal of attention to uncertainty and incomplete information; to so-called "agency theory," which takes account of the institutional framework within which decisions are made; and to game theory, which seeks to deal with interindividual and intergroup processes in which there is partial conflict of interest. Economists and political scientists are also increasingly buttressing the empirical foundations of their field by studying individual choice behavior directly and by studying behavior in experimentally constructed markets and simulated political structures.

The following pages contain a fuller outline of current knowledge about decision making and problem solving and a brief review of current research directions in these fields as well as some of the principal research opportunities.

DECISION MAKING

SEU Theory

The development of SEU theory was a major intellectual achievement of the first half of this century. It gave for the first time a formally axiomatized statement of what it would mean for an agent to behave in a consistent, rational manner. It assumed that a decision maker possessed a utility function (an ordering by preference among all the possible outcomes of choice), that all the alternatives among which choice could be made were known, and that the consequences of choosing each alternative could be ascertained (or, in the version of the theory that treats of choice under uncertainty, it assumed that a subjective or objective probability distribution of consequences was associated with each alternative). By admitting subjectively assigned probabilities, SEU theory opened the way to fusing subjective opinions with objective data, an approach that can also be used in man-machine decision-making systems. In the probabilistic version of the theory, Bayes's rule prescribes how people should take account of new information and how they should respond to incomplete information.

The assumptions of SEU theory are very strong, permitting correspondingly strong inferences to be made from them. Although the assumptions cannot be satisfied even remotely for most complex situations in the real world, they may be satisfied approximately in some microcosms—problem situations that can be isolated from the world's complexity and dealt with independently. For example, the manager of a commercial cattle-feeding operation might isolate the problem of finding the least expensive mix of feeds available in the market that would meet all the nutritional requirements of his cattle. The computational tool of linear programming, which is a powerful method for maximizing goal achievement or minimizing costs while satisfying all kinds of side conditions (in this case, the nutritional requirements), can provide the manager with an optimal feed mix—optimal within the limits of approximation of his model to real-world conditions. Linear programming and related operations research techniques are now used widely to make decisions whenever a situation that reasonably fits their assumptions can be carved out of its complex surround. These techniques have been especially valuable aids to middle management in dealing with relatively well-structured decision problems.

Most of the tools of modern operations research—not only linear programming, but also integer programming, queuing theory, decision trees, and other widely used techniques—use the assumptions of SEU theory. They assume that what is desired is to maximize the achievement of some goal, under specified constraints and assuming that all

alternatives and consequences (or their probability distributions) are known. These tools have proven their usefulness in a wide variety of applications.

The Limits of Rationality

Operations research tools have also underscored dramatically the limits of SEU theory in dealing with complexity. For example, present and prospective computers are not even powerful enough to provide exact solutions for the problems of optimal scheduling and routing of jobs through a typical factory that manufactures a variety of products using many different tools and machines. And the mere thought of using these computational techniques to determine an optimal national policy for energy production or an optimal economic policy reveals their limits.

Computational complexity is not the only factor that limits the literal application of SEU theory. The theory also makes enormous demands on information. For the utility function, the range of available alternatives and the consequences following from each alternative must all be known. Increasingly, research is being directed at decision making that takes realistic account of the compromises and approximations that must be made in order to fit real-world problems to the informational and computational limits of people and computers, as well as to the inconsistencies in their values and perceptions. The study of actual decision processes (for example, the strategies used by corporations to make their investments) reveals massive and unavoidable departures from the framework of SEU theory. The sections that follow describe some of the things that have been learned about choice under various conditions of incomplete information, limited computing power, inconsistency, and institutional constraints on alternatives. Game theory, agency theory, choice under uncertainty, and the theory of markets are a few of the directions of this research, with the aims both of constructing prescriptive theories of broader application and of providing more realistic descriptions and explanations of actual decision making within U.S. economic and political institutions.

Limited Rationality in Economic Theory

Although the limits of human rationality were stressed by some researchers in the 1950s, only recently has there been extensive activity in the field of economics aimed at developing theories that assume less than fully rational choice on the part of business firm managers and other economic agents. The newer theoretical research undertakes to answer such questions as the following:

- Are market equilibria altered by the departures of actual choice behavior from the behavior of fully rational agents predicted by SEU theory?
- Under what circumstances do the processes of competition “police” markets in such a way as to cancel out the effects of the departures from full rationality?
- In what ways are the choices made by boundedly rational agents different from those made by fully rational agents?

Theories of the firm that assume managers are aiming at “satisfactory” profits or that their concern is to maintain the firm's share of market in the industry make quite different predictions about economic equilibrium than those derived from the assumption of profit maximization. Moreover, the classical theory of the firm cannot explain why economic activity is sometimes organized around large business firms and sometimes around contractual networks of individuals or smaller organizations. New theories that take account of differential access of economic agents to information, combined with differences in self-interest, are able to account for these important phenomena, as well as provide explanations for

the many forms of contracts that are used in business. Incompleteness and asymmetry of information have been shown to be essential for explaining how individuals and business firms decide when to face uncertainty by insuring, when by hedging, and when by assuming the risk.

Most current work in this domain still assumes that economic agents seek to maximize utility, but within limits posed by the incompleteness and uncertainty of the information available to them. An important potential area of research is to discover how choices will be changed if there are other departures from the axioms of rational choice—for example, substituting goals of reaching specified aspiration levels (satisficing) for goals of maximizing.

Applying the new assumptions about choice to economics leads to new empirically supported theories about decision making over time. The classical theory of perfect rationality leaves no room for regrets, second thoughts, or "weakness of will." It cannot explain why many individuals enroll in Christmas savings plans, which earn interest well below the market rate. More generally, it does not lead to correct conclusions about the important social issues of saving and conservation. The effect of pensions and social security on personal saving has been a controversial issue in economics. The standard economic model predicts that an increase in required pension saving will reduce other saving dollar for dollar; behavioral theories, on the other hand, predict a much smaller offset. The empirical evidence indicates that the offset is indeed very small. Another empirical finding is that the method of payment of wages and salaries affects the saving rate. For example, annual bonuses produce a higher saving rate than the same amount of income paid in monthly salaries. This finding implies that saving rates can be influenced by the way compensation is framed.

If individuals fail to discount properly for the passage of time, their decisions will not be optimal. For example, air conditioners vary greatly in their energy efficiency; the more efficient models cost more initially but save money over the long run through lower energy consumption. It has been found that consumers, on average, choose air conditioners that imply a discount rate of 25 percent or more per year, much higher than the rates of interest that prevailed at the time of the study.

As recently as five years ago, the evidence was thought to be unassailable that markets like the New York Stock Exchange work efficiently—that prices reflect all available information at any given moment in time, so that stock price movements resemble a random walk and contain no systematic information that could be exploited for profit. Recently, however, substantial departures from the behavior predicted by the efficient-market hypothesis have been detected. For example, small firms appear to earn inexplicably high returns on the market prices of their stock, while firms that have very low price-earnings ratios and firms that have lost much of their market value in the recent past also earn abnormally high returns. All of these results are consistent with the empirical finding that decision makers often overreact to new information, in violation of Bayes's rule. In the same way, it has been found that stock prices are excessively volatile—that they fluctuate up and down more rapidly and violently than they would if the market were efficient.

There has also been a long-standing puzzle as to why firms pay dividends. Considering that dividends are taxed at a higher rate than capital gains, taxpaying investors should prefer, under the assumptions of perfect rationality, that their firms reinvest earnings or repurchase shares instead of paying dividends. (The investors could simply sell some of their appreciated shares to obtain the income they require.) The solution to this puzzle also requires models of investors that take account of limits on rationality.

The Theory of Games

In economic, political, and other social situations in which there is actual or potential conflict of interest, especially if it is combined with incomplete information, SEU theory faces special difficulties. In markets in which there are many competitors (e.g., the wheat market), each buyer or seller can accept the market price as a "given" that will not be affected materially by the actions of any single individual. Under these conditions, SEU theory makes unambiguous predictions of behavior. However, when a market has only a few suppliers—say, for example, two—matters are quite different. In this case, what it is rational to do depends on what one's competitor is going to do, and vice versa. Each supplier may try to outwit the other. What then is the rational decision?

The most ambitious attempt to answer questions of this kind was the theory of games, developed by von Neumann and Morgenstern and published in its full form in 1994. But the answers provided by the theory of games are sometimes very puzzling and ambiguous. In many situations, no single course of action dominates all the others; instead, a whole set of possible solutions are all equally consistent with the postulates of rationality.

One game that has been studied extensively, both theoretically and empirically, is the Prisoner's Dilemma. In this game between two players, each has a choice between two actions, one trustful of the other player, the other mistrustful or exploitative. If both players choose the trustful alternative, both receive small rewards. If both choose the exploitative alternative, both are punished. If one chooses the trustful alternative and the other the exploitative alternative, the former is punished much more severely than in the previous case, while the latter receives a substantial reward. If the other player's choice is fixed but unknown, it is advantageous for a player to choose the exploitative alternative, for this will give him the best outcome in either case. But if both adopt this reasoning, they will both be punished, whereas they could both receive rewards if they agreed upon the trustful choice (and did not welch on the agreement).

The terms of the game have an unsettling resemblance to certain situations in the relations between nations or between a company and the employees' union. The resemblance becomes stronger if one imagines the game as being played repeatedly. Analyses of "rational" behavior under assumptions of intended utility maximization support the conclusion that the players will (ought to?) always make the mistrustful choice. Nevertheless, in laboratory experiments with the game, it is often found that players (even those who are expert in game theory) adopt a "tit-for-tat" strategy. That is, each plays the trustful, cooperative strategy as long as his or her partner does the same. If the partner exploits the player on a particular trial, the player then plays the exploitative strategy on the next trial and continues to do so until the partner switches back to the trustful strategy. Under these conditions, the game frequently stabilizes with the players pursuing the mutually trustful strategy and receiving the rewards.

With these empirical findings in hand, theorists have recently sought and found some of the conditions for attaining this kind of benign stability. It occurs, for example, if the players set aspirations for a satisfactory reward rather than seeking the maximum reward. This result is consistent with the finding that in many situations, as in the Prisoner's Dilemma game, people appear to satisfice rather than attempting to optimize.

The Prisoner's Dilemma game illustrates an important point that is beginning to be appreciated by those who do research on decision making. There are so many ways in which actual human behavior can depart from the SEU assumptions that theorists seeking to account for behavior are confronted with an embarrassment of riches.

To choose among the many alternative models that could account for the anomalies of choice, extensive empirical research is called for—to see how people do make their choices, what beliefs guide them, what information they have available, and what part of that information they take into account and what part they ignore. In a world of limited rationality, economics and the other decision sciences must closely examine the actual limits on rationality in order to make accurate predictions and to provide sound advice on public policy.

Empirical Studies of Choice Under Uncertainty

During the past 10 years, empirical studies of human choices in which uncertainty, inconsistency, and incomplete information are present have produced a rich collection of findings which only now are beginning to be organized under broad generalizations. Here are a few examples. When people are given information about the probabilities of certain events (e.g., how many lawyers and how many engineers are in a population that is being sampled), and then are given some additional information as to which of the events has occurred (which person has been sampled from the population), they tend to ignore the prior probabilities in favor of incomplete or even quite irrelevant information about the individual event. Thus, if they are told that 70 percent of the population are lawyers, and if they are then given a noncommittal description of a person (one that could equally well fit a lawyer or an engineer), half the time they will predict that the person is a lawyer and half the time that he is an engineer—even though the laws of probability dictate that the best forecast is always to predict that the person is a lawyer.

People commonly misjudge probabilities in many other ways. Asked to estimate the probability that 60 percent or more of the babies born in a hospital during a given week are male, they ignore information about the total number of births, although it is evident that the probability of a departure of this magnitude from the expected value of 50 percent is smaller if the total number of births is larger (the standard error of a percentage varies inversely with the square root of the population size).

There are situations in which people assess the frequency of a class by the ease with which instances can be brought to mind. In one experiment, subjects heard a list of names of persons of both sexes and were later asked to judge whether there were more names of men or women on the list. In lists presented to some subjects, the men were more famous than the women; in other lists, the women were more famous than the men. For all lists, subjects judged that the sex that had the more famous personalities was the more numerous.

The way in which an uncertain possibility is presented may have a substantial effect on how people respond to it. When asked whether they would choose surgery in a hypothetical medical emergency, many more people said that they would when the chance of survival was given as 80 percent than when the chance of death was given as 20 percent.

On the basis of these studies, some of the general heuristics, or rules of thumb, that people use in making judgments have been compiled—heuristics that produce biases toward classifying situations according to their representativeness, or toward judging frequencies according to the availability of examples in memory, or toward interpretations warped by the way in which a problem has been framed. These findings have important implications for public policy. A recent example is the lobbying effort of the credit card industry to have differentials between cash and credit prices labeled “cash discounts” rather than “credit surcharges.” The research findings raise questions about how to phrase cigarette warning labels or

frame truth-in-lending laws and informed-consent laws.

Methods of Empirical Research

Finding the underlying bases of human choice behavior is difficult. People cannot always, or perhaps even usually, provide veridical accounts of how they make up their minds, especially when there is uncertainty. In many cases, they can predict how they will behave (pre-election polls of voting intentions have been reasonably accurate when carefully taken), but the reasons people give for their choices can often be shown to be rationalizations and not closely related to their real motives.

Students of choice behavior have steadily improved their research methods. They question respondents about specific situations, rather than asking for generalizations. They are sensitive to the dependence of answers on the exact forms of the questions. They are aware that behavior in an experimental situation may be different from behavior in real life, and they attempt to provide experimental settings and motivations that are as realistic as possible. Using thinking-aloud protocols and other approaches, they try to track the choice behavior step by step, instead of relying just on information about outcomes or querying respondents retrospectively about their choice processes.

Perhaps the most common method of empirical research in this field is still to ask people to respond to a series of questions. But data obtained by this method are being supplemented by data obtained from carefully designed laboratory experiments and from observations of actual choice behavior (for example, the behavior of customers in supermarkets). In an experimental study of choice, subjects may trade in an actual market with real (if modest) monetary rewards and penalties. Research experience has also demonstrated the feasibility of making direct observations, over substantial periods of time, of the decision-making processes in business and governmental organizations—for example, observations of the procedures that corporations use in making new investments in plant and equipment. Confidence in the empirical findings that have been accumulating over the past several decades is enhanced by the general consistency that is observed among the data obtained from quite different settings using different research methods.

There still remains the enormous and challenging task of putting together these findings into an empirically founded theory of decision making. With the growing availability of data, the theory-building enterprise is receiving much better guidance from the facts than it did in the past. As a result, we can expect it to become correspondingly more effective in arriving at realistic models of behavior.

PROBLEM SOLVING

The theory of choice has its roots mainly in economics, statistics, and operations research and only recently has received much attention from psychologists; the theory of problem solving has a very different history. Problem solving was initially studied principally by psychologists, and more recently by researchers in artificial intelligence. It has received rather scant attention from economists.

Contemporary Problem-Solving Theory

Human problem solving is usually studied in laboratory settings, using problems that can be solved in relatively short periods of time (seldom more than an hour), and often seeking a maximum density of data about the solution process by asking subjects to think aloud while they work. The thinking-aloud technique, at first viewed with suspicion by behaviorists as subjective and "introspective," has received such careful

methodological attention in recent years that it can now be used dependably to obtain data about subjects' behaviors in a wide range of settings.

The laboratory study of problem solving has been supplemented by field studies of professionals solving real-world problems—for example, physicians making diagnoses and chess grandmasters analyzing game positions, and, as noted earlier, even business corporations making investment decisions. Currently, historical records, including laboratory notebooks of scientists, are also being used to study problem-solving processes in scientific discovery. Although such records are far less "dense" than laboratory protocols, they sometimes permit the course of discovery to be traced in considerable detail. Laboratory notebooks of scientists as distinguished as Charles Darwin, Michael Faraday, Antoine-Laurent Lavoisier, and Hans Krebs have been used successfully in such research.

From empirical studies, a description can now be given of the problem-solving process that holds for a rather wide range of activities. First, problem solving generally proceeds by selective search through large sets of possibilities, using rules of thumb (heuristics) to guide the search. Because the possibilities in realistic problem situations are generally multitudinous, trial-and-error search would simply not work; the search must be highly selective. Chess grandmasters seldom examine more than a hundred of the vast number of possible scenarios that confront them, and similar small numbers of searches are observed in other kinds of problem-solving search.

One of the procedures often used to guide search is "hill climbing," using some measure of approach to the goal to determine where it is most profitable to look next. Another, and more powerful, common procedure is means-ends analysis. In means-end analysis, the problem solver compares the present situation with the goal, detects a difference between them, and then searches memory for actions that are likely to reduce the difference. Thus, if the difference is a 50-mile distance from the goal, the problem solver will retrieve from memory knowledge about autos, carts, bicycles, and other means of transport; walking and flying will probably be discarded as inappropriate for that distance.

The third thing that has been learned about problem solving—especially when the solver is an expert—is that it relies on large amounts of information that are stored in memory and that are retrievable whenever the solver recognizes cues signaling its relevance. Thus, the expert knowledge of a diagnostician is evoked by the symptoms presented by the patient; this knowledge leads to the recollection of what additional information is needed to discriminate among alternative diseases and, finally, to the diagnosis.

In a few cases, it has been possible to estimate how many patterns an expert must be able to recognize in order to gain access to the relevant knowledge stored in memory. A chess master must be able to recognize about 50,000 different configurations of chess pieces that occur frequently in the course of chess games. A medical diagnostician must be able to recognize tens of thousands of configurations of symptoms; a botanist or zoologist specializing in taxonomy, tens or hundreds of thousands of features of specimens that define their species. For comparison, college graduates typically have vocabularies in their native languages of 50,000 to 200,000 words. (However, these numbers are very small in comparison with the real-world situations the expert faces: there are perhaps 10^{120} branches in the game tree of chess, a game played with only six kinds of pieces on an 8×8 board.)

One of the accomplishments of the contemporary theory of problem solving has been to provide an explanation for the phenomena of intuition and judgment frequently seen in experts' behavior. The store of expert knowledge, "indexed" by the rec

ognition cues that make it accessible and combined with some basic inferential capabilities (perhaps in the form of means-ends analysis), accounts for the ability of experts to find satisfactory solutions for difficult problems, and sometimes to find them almost instantaneously. The expert's "intuition" and "judgment" derive from this capability for rapid recognition linked to a large store of knowledge. When immediate intuition fails to yield a problem solution or when a prospective solution needs to be evaluated, the expert falls back on the slower processes of analysis and inference.

Expert Systems in Artificial Intelligence

Over the past 30 years, there has been close teamwork between research in psychology and research in computer science aimed at developing intelligent programs. Artificial intelligence (AI) research has both borrowed from and contributed to research on human problem solving. Today, artificial intelligence is beginning to produce systems, applied to a variety of tasks, that can solve difficult problems at the level of professionally trained humans. These AI programs are usually called expert systems. A description of a typical expert system would resemble closely the description given above of typical human problem solving; the differences between the two would be differences in degree, not in kind. An AI expert system, relying on the speed of computers and their ability to retain large bodies of transient information in memory, will generally use "brute force"—sheer computational speed and power—more freely than a human expert can. A human expert, in compensation, will generally have a richer set of heuristics to guide search and a larger vocabulary of recognizable patterns. To the observer, the computer's process will appear the more systematic and even compulsive, the human's the more intuitive. But these are quantitative, not qualitative, differences.

The number of tasks for which expert systems have been built is increasing rapidly. One is medical diagnosis (two examples are the CADUCEUS and MYCIN programs). Others are automatic design of electric motors, generators, and transformers (which predates by a decade the invention of the term "expert systems"), the configuration of computer systems from customer specifications, and the automatic generation of reaction paths for the synthesis of organic molecules. All of these (and others) are either being used currently in professional or industrial practice or at least have reached a level at which they can produce a professionally acceptable product.

Expert systems are generally constructed in close consultation with the people who are experts in the task domain. Using standard techniques of observation and interrogation, the heuristics that the human expert uses, implicitly and often unconsciously, to perform the task are gradually reduced, made explicit, and incorporated in program structures. Although a great deal has been learned about how to do this, improving techniques for designing expert systems is an important current direction of research. It is especially important because expert systems, once built, cannot remain static but must be modifiable to incorporate new knowledge as it becomes available.

Dealing with Ill-Structured Problems

In the 1950s and 1960s, research on problem solving focused on clearly structured puzzle-like problems that were easily brought into the psychological laboratory and that were within the range of computer programming sophistication at that time. Computer programs were written to discover proofs for theorems in Euclidean geometry or to solve the puzzle of transporting missionaries and cannibals across a river. Choosing chess moves was perhaps the most complex task that received attention in the early years of cognitive science and AI.

As understanding grew of the methods needed to handle these relatively simple tasks, research aspirations rose. The next main target, in the 1960s and 1970s, was to find methods for solving problems that involved large bodies of semantic information. Medical diagnosis and interpreting mass spectrogram data are examples of the kinds of tasks that were investigated during this period and for which a good level of understanding was achieved. They are tasks that, for all of the knowledge they call upon, are still well structured, with clear-cut goals and constraints.

The current research target is to gain an understanding of problem-solving tasks when the goals themselves are complex and sometimes ill defined, and when the very nature of the problem is successively transformed in the course of exploration. To the extent that a problem has these characteristics, it is usually called ill structured. Because ambiguous goals and shifting problem formulations are typical characteristics of problems of design, the work of architects offers a good example of what is involved in solving ill-structured problems. An architect begins with some very general specifications of what is wanted by a client. The initial goals are modified and substantially elaborated as the architect proceeds with the task. Initial design ideas, recorded in drawings and diagrams, themselves suggest new criteria, new possibilities, and new requirements. Throughout the whole process of design, the emerging conception provides continual feedback that reminds the architect of additional considerations that need to be taken into account.

With the current state of the art, it is just beginning to be possible to construct programs that simulate this kind of flexible problem-solving process. What is called for is an expert system whose expertise includes substantial knowledge about design criteria as well as knowledge about the means for satisfying those criteria. Both kinds of knowledge are evoked in the course of the design activity by the usual recognition processes, and the evocation of design criteria and constraints continually modifies and remolds the problem that the design system is addressing. The large data bases that can now be constructed to aid in the management of architectural and construction projects provide a framework into which AI tools, fashioned along these lines, can be incorporated.

Most corporate strategy problems and governmental policy problems are at least as ill structured as problems of architectural or engineering design. The tools now being forged for aiding architectural design will provide a basis for building tools that can aid in formulating, assessing, and monitoring public energy or environmental policies, or in guiding corporate product and investment strategies.

Setting the Agenda and Representing a Problem

The very first steps in the problem-solving process are the least understood. What brings (and should bring) problems to the head of the agenda? And when a problem is identified, how can it be represented in a way that facilitates its solution?

The task of setting an agenda is of utmost importance because both individual human beings and human institutions have limited capacities for dealing with many tasks simultaneously. While some problems are receiving full attention, others are neglected. When new problems come thick and fast, “fire fighting” replaces planning and deliberation. The facts of limited attention span, both for individuals and for institutions like the Congress, are well known. However, relatively little has been accomplished toward analyzing or designing effective agenda-setting systems. A beginning could be made by the study of “alerting” organizations like the Office of Technology Assessment or military and foreign affairs intelligence agencies. Because the research

and development function in industry is also in considerable part a task of monitoring current and prospective technological advances, it could also be studied profitably from this standpoint.

The way in which problems are represented has much to do with the quality of the solutions that are found. The task of designing highways or dams takes on an entirely new aspect if human responses to a changed environment are taken into account. (New transportation routes cause people to move their homes, and people show a considerable propensity to move into zones that are subject to flooding when partial protections are erected.) Very different social welfare policies are usually proposed in response to the problem of providing incentives for economic independence than are proposed in response to the problem of taking care of the needy. Early management information systems were designed on the assumption that information was the scarce resource; today, because designers recognize that the scarce resource is managerial attention, a new framework produces quite different designs.

The representation or "framing" of problems is even less well understood than agenda setting. Today's expert systems make use of problem representations that already exist. But major advances in human knowledge frequently derive from new ways of thinking about problems. A large part of the history of physics in nineteenth-century England can be written in terms of the shift from action-at-a-distance representations to the field representations that were developed by the applied mathematicians at Cambridge.

Today, developments in computer-aided design (CAD) present new opportunities to provide human designers with computer-generated representations of their problems. Effective use of these capabilities requires us to understand better how people extract information from diagrams and other displays and how displays can enhance human performance in design tasks. Research on representations is fundamental to the progress of CAD.

Computation as Problem Solving

Nothing has been said so far about the radical changes that have been brought about in problem solving over most of the domains of science and engineering by the standard uses of computers as computational devices. Although a few examples come to mind in which artificial intelligence has contributed to these developments, they have mainly been brought about by research in the individual sciences themselves, combined with work in numerical analysis.

Whatever their origins, the massive computational applications of computers are changing the conduct of science in numerous ways. There are new specialties emerging such as "computational physics" and "computational chemistry." Computation—that is to say, problem solving—becomes an object of explicit concern to scientists, side by side with the substance of the science itself. Out of this new awareness of the computational component of scientific inquiry is arising an increasing interaction among computational specialists in the various sciences and scientists concerned with cognition and AI. This interaction extends well beyond the traditional area of numerical analysis, or even the newer subject of computational complexity, into the heart of the theory of problem solving.

Physicists seeking to handle the great mass of bubble-chamber data produced by their instruments began, as early as the 1960s, to look to AI for pattern recognition methods as a basis for automating the analysis of their data. The construction of expert systems to interpret mass spectrogram data and of other systems to design synthesis paths for chemical reactions are other examples of problem solving in science, as are programs to aid in matching sequences of nucleic acids in DNA

and RNA and amino acid sequences in proteins.

Theories of human problem solving and learning are also beginning to attract new attention within the scientific community as a basis for improving science teaching. Each advance in the understanding of problem solving and learning processes provides new insights about the ways in which a learner must store and index new knowledge and procedures if they are to be useful for solving problems. Research on these topics is also generating new ideas about how effective learning takes place—for example, how students can learn by examining and analyzing worked-out examples.

Extensions of Theory

Opportunities for advancing our understanding of decision making and problem solving are not limited to the topics dealt with above, and in this section, just a few indications of additional promising directions for research are presented.

Decision Making over Time

The time dimension is especially trouble-some in decision making. Economics has long used the notion of time discounting and interest rates to compare present with future consequences of decisions, but as noted above, research on actual decision making shows that people frequently are inconsistent in their choices between present and future. Although time discounting is a powerful idea, it requires fixing appropriate discount rates for individual, and especially social, decisions. Additional problems arise because human tastes and priorities change over time. Classical SEU theory assumes a fixed, consistent utility function, which does not easily accommodate changes in taste. At the other extreme, theories postulating a limited attention span do not have ready ways of ensuring consistency of choice over time.

Aggregation

In applying our knowledge of decision making and problem solving to society-wide, or even organization-wide, phenomena, the problem of aggregation must be solved; that is, ways must be found to extrapolate from theories of individual decision processes to the net effects on the whole economy, polity, and society. Because of the wide variety of ways in which any given decision task can be approached, it is unrealistic to postulate a "representative firm" or an "economic man," and to simply lump together the behaviors of large numbers of supposedly identical individuals. Solving the aggregation problem becomes more important as more of the empirical research effort is directed toward studying behavior at a detailed, microscopic level.

Organizations

Related to aggregation is the question of how decision making and problem solving change when attention turns from the behavior of isolated individuals to the behavior of these same individuals operating as members of organizations or other groups. When people assume organizational positions, they adapt their goals and values to their responsibilities. Moreover, their decisions are influenced substantially by the patterns of information flow and other communications among the various organization units.

Organizations sometimes display sophisticated capabilities far beyond the understanding of single individuals. They sometimes make enormous blunders or find themselves incapable of acting. Organizational performance is highly sensitive to the quality of the routines or "performance programs" that govern behavior and to the adaptability of these routines in the face of a changing environment. In particular, the "peripheral vision" of a complex organization is limited, so that responses to novelty

in the environment may be made in inappropriate and quasi-automatic ways that cause major failure.

Theory development, formal modeling, laboratory experiments, and analysis of historical cases are all going forward in this important area of inquiry. Although the decision-making processes of organizations have been studied in the field on a limited scale, a great many more such intensive studies will be needed before the full range of techniques used by organizations to make their decisions is understood, and before the strengths and weaknesses of these techniques are grasped.

Learning

Until quite recently, most research in cognitive science and artificial intelligence had been aimed at understanding how intelligent systems perform their work. Only in the past five years has attention begun to turn to the question of how systems become intelligent—how they learn. A number of promising hypotheses about learning mechanisms are currently being explored. One is the so-called connexionist hypothesis, which postulates networks that learn by changing the strengths of their interconnections in response to feedback. Another learning mechanism that is being investigated is the adaptive production system, a computer program that learns by generating new instructions that are simply annexed to the existing program. Some success has been achieved in constructing adaptive production systems that can learn to solve equations in algebra and to do other tasks at comparable levels of difficulty.

Learning is of particular importance for successful adaptation to an environment that is changing rapidly. Because that is exactly the environment of the 1980s, the trend toward broadening research on decision making to include learning and adaptation is welcome.

This section has by no means exhausted the areas in which exciting and important research can be launched to deepen understanding of decision making and problem solving. But perhaps the examples that have been provided are sufficient to convey the promise and significance of this field of inquiry today.

CURRENT RESEARCH PROGRAMS

Most of the current research on decision making and problem solving is carried on in universities, frequently with the support of government funding agencies and private foundations. Some research is done by consulting firms in connection with their development and application of the tools of operations research, artificial intelligence, and systems modeling. In some cases, government agencies and corporations have supported the development of planning models to aid them in their policy planning—for example, corporate strategic planning for investments and markets and government planning of environmental and energy policies. There is an increasing number of cases in which research scientists are devoting substantial attention to improving the problem-solving and decision-making tools in their disciplines, as we noted in the examples of automation of the processing of bubble-chamber tracks and of the interpretation of mass spectrogram data.

To use a generous estimate, support for basic research in the areas described in this document is probably at the level of tens of millions of dollars per year, and almost certainly, it is not as much as $100 million. The principal costs are for research personnel and computing equipment, the former being considerably larger.

Because of the interdisciplinary character of the research domain, federal research support comes from a number of different agencies, and it is not easy to assess the total picture. Within the National Science Foundation (NSF), the grants of the decision and

management sciences, political science and the economics programs in the Social Sciences Division are to a considerable extent devoted to projects in this domain. Smaller amounts of support come from the memory and cognitive processes program in the Division of Behavioral and Neural Sciences, and perhaps from other programs. The "software" component of the new NSF Directorate of Computer Science and Engineering contains programs that have also provided important support to the study of decision making and problem solving.

The Office of Naval Research has, over the years, supported a wide range of studies of decision making, including important early support for operations research. The main source of funding for research in AI has been the Defense Advanced Research Projects Agency (DARPA) in the Department of Defense; important support for research on applications of AI to medicine has been provided by the National Institutes of Health.

Relevant economics research is also funded by other federal agencies, including the Treasury Department, the Bureau of Labor Statistics, and the Federal Reserve Board. In recent years, basic studies of decision making have received only relatively minor support from these sources, but because of the relevance of the research to their missions, they could become major sponsors.

Although a number of projects have been and are funded by private foundations, there appears to be at present no foundation for which decision making and problem solving are a major focus of interest.

In sum, the pattern of support for research in this field shows a healthy diversity but no agency with a clear lead responsibility, unless it be the rather modestly funded program in decision and management sciences at NSF. Perhaps the largest scale of support has been provided by DARPA, where decision making and problem solving are only components within the larger area of artificial intelligence and certainly not highly visible research targets.

The character of the funding requirements in this domain is much the same as in other fields of research. A rather intensive use of computational facilities is typical of most, but not all, of the research. And because the field is gaining new recognition and growing rapidly, there are special needs for the support of graduate students and postdoctoral training. In the computing-intensive part of the domain, desirable research funding per principal investigator might average $250,000 per year; in empirical research involving field studies and largescale experiments, a similar amount; and in other areas of theory and laboratory experimentation, somewhat less.

RESEARCH OPPORTUNITIES: SUMMARY

The study of decision making and problem solving has attracted much attention through most of this century. By the end of World War II, a powerful prescriptive theory of rationality, the theory of subjective expected utility (SEU), had taken form; it was followed by the theory of games. The past 40 years have seen widespread applications of these theories in economics, operations research, and statistics, and, through these disciplines, to decision making in business and government.

The main limitations of SEU theory and the developments based on it are its relative neglect of the limits of human (and computer) problem-solving capabilities in the face of real-world complexity. Recognition of these limitations has produced an increasing volume of empirical research aimed at discovering how humans cope with complexity and reconcile it with their bounded computational powers. Recognition that human rationality is limited occasions no surprise. What is surprising are some of the forms these limits take and the kinds of departures from the behavior predicted by the SEU model that have been observed. Extending empirical knowledge of actual hu

man cognitive processes and of techniques for dealing with complexity continues to be a research goal of very high priority. Such empirical knowledge is needed both to build valid theories of how the U.S. society and economy operate and to build prescriptive tools for decision making that are compatible with existing computational capabilities.

The complementary fields of cognitive psychology and artificial intelligence have produced in the past 30 years a fairly well-developed theory of problem solving that lends itself well to computer simulation, both for purposes of testing its empirical validity and for augmenting human problem-solving capacities by the construction of expert systems. Problem-solving research today is being extended into the domain of ill-structured problems and applied to the task of formulating problem representations. The processes for setting the problem agenda, which are still very little explored, deserve more research attention.

The growing importance of computational techniques in all of the sciences has attracted new attention to numerical analysis and to the topic of computational complexity. The need to use heuristic as well as rigorous methods for analyzing very complex domains is beginning to bring about a wide interest, in various sciences, in the possible application of problem-solving theories to computation.

Opportunities abound for productive research in decision making and problem solving. A few of the directions of research that look especially promising and significant follow:

- A substantially enlarged program of empirical studies, involving direct observation of behavior at the level of the individual and the organization, and including both laboratory and field experiments, will be essential in sifting the wheat from the chaff in the large body of theory that now exists and in giving direction to the development of new theory.
- Expanded research on expert systems will require extensive empirical study of expert behavior and will provide a setting for basic research on how ill-structured problems are, and can be, solved.
- Decision making in organizational settings, which is much less well understood than individual decision making and problem solving, can be studied with great profit using already established methods of inquiry, especially through intensive long-range studies within individual organizations.
- The resolution of conflicts of values (individual and group) and of inconsistencies in belief will continue to be highly productive directions of inquiry, addressed to issues of great importance to society.
- Setting agendas and framing problems are two related but poorly understood processes that require special research attention and that now seem open to attack.

These five areas are examples of especially promising research opportunities drawn from the much larger set that are described or hinted at in this report.

The tools for decision making developed by previous research have already found extensive application in business and government organizations. A number of such applications have been mentioned in this report, but they so pervade organizations, especially at the middle management and professional levels, that people are often unaware of their origins.

Although the research domain of decision making and problem solving is alive and well today, the resources devoted to that research are modest in scale (of the order of tens of millions rather than hundreds of millions of dollars). They are not commensurate with either the identified research opportunities or the human resources available for exploiting them. The prospect of throwing new light on the ancient problem of mind and the prospect of enhancing the powers of mind with new computational

tools are attracting substantial numbers of first-rate young scientists. Research progress is not limited either by lack of excellent research problems or by lack of human talent eager to get on with the job.

Gaining a better understanding of how problems can be solved and decisions made is essential to our national goal of increasing productivity. The first industrial revolution showed us how to do most of the world's heavy work with the energy of machines instead of human muscle. The new industrial revolution is showing us how much of the work of human thinking can be done by and in cooperation with intelligent machines. Human minds with computers to aid them are our principal productive resource. Understanding how that resource operates is the main road open to us for becoming a more productive society and a society able to deal with the many complex problems in the world today.

Report of the Research Briefing Panel on Protein Structure and Biological Function

Research Briefing Panel on Protein Structure and Biological Function

Frederic Richards *(Chairman),* Yale University, New Haven, Conn.
Robert Baldwin, Stanford University Medical Center, Palo Alto, Calif.
Gerald R. Galluppi, Monsanto Company, St. Louis, Mo.
Robert Griffin, Massachusetts Institute of Technology, Cambridge, Mass.
Emil Thomas Kaiser, Rockefeller University, New York, N.Y.
Brain Matthews, University of Oregon, Eugene, Oreg.
J. Andrew McCammon, University of Houston-University Park, Houston, Tex.
Alfred Redfield, Brandeis University, Waltham, Mass.
Brain Reid, University of Washington, Seattle, Wash.
Robert Sauer, Massachusetts Institute of Technology, Cambridge, Mass.
Alan Schechter, National Institutes of Health, Bethesda, Md.
Paul Sigler, University of Chicago, Chicago, Ill.
Peter von Hippel, University of Oregon, Eugene, Oreg.
Don Wiley, Harvard University, Cambridge, Mass.

Staff

Barbara Filner, *Director,* Division of Health Sciences Policy, Institute of Medicine
Naomi Hudson, *Administrative Secretary*
Allan R. Hoffman, *Executive Director,* Committee on Science, Engineering, and Public Policy

Report of the Research Briefing Panel on Protein Structure and Biological Function

INTRODUCTION

Proteins are involved in every biological function. As enzymes, they catalyze the chemical reactions of cells. As hormones and growth factors, they regulate the development of cells and coordinate the functions of distant organs in the body. In various filamentous forms, they control the shape of cells and the dramatic alterations that occur during cell division. In muscle, proteins change chemical energy into mechanical energy and cause movement. As components of membranes, they control the traffic of molecules and information among the various cellular compartments. Hemoglobin, a protein in blood, is specifically designed to transport oxygen between organs; other blood proteins, such as clotting factors and circulating antibodies, act as defenses against trauma and infection. In plants, a highly organized collection of membrane proteins is involved in the complex process of photosynthesis, without which there would be no higher animal forms.

Proteins are polymer molecules composed of amino acids that are connected by links known as peptide bonds. An individual protein molecule may contain hundreds or thousands of amino acids arranged in one, or several, polypeptide chains. Each chain folds into a particular three-dimensional configuration that is essential for its highly specific biological function. In this report, we focus on the experimental and theoretical investigation of the three-dimensional structure of proteins, at the level of resolution of individual atoms.

The study of protein structures, frequently referred to as structural biology, is in a period of great excitement brought about by developments in several fields. Many proteins of special interest are now available in unprecedented amounts. The chemical synthesis of polypeptide chains of increasing size has steadily improved over the past 20 years, and recently a chain length of over 100 amino acids was achieved. In the past 5 years, the biological synthesis of proteins, through cloning, has been reduced to standard practice, not only in the laboratory but also on a commercial scale. These complementary procedures supply proteins of defined sequence in substantial amounts. At the same time, spectacular advances have been made in x-ray diffraction and nuclear magnetic resonance spectroscopy, two techniques for determining structure. New pro

cedures and refined equipment have expanded the range of application and the size of the proteins that can be studied. Chemical theory also has been developing at a rapid rate, especially those branches related to polymers. Furthermore, with the distribution of faster, smaller, and less expensive computer hardware, there has been constant improvement in access to more advanced computing capability, a factor that has played an important role in theoretical and experimental studies. The coming together of these separate developments makes structural biology ready for an explosive increase in the determination and understanding of high-resolution structures of proteins and protein complexes.

The new surge of structural information will dramatically improve our understanding of the processes of biological control and will guide the design of proteins or other products that will be developed either to cause purposeful malfunctions (e.g., insecticides) or to correct natural malfunctions (e.g., to improve human health). Guided by the three-dimensional structure, changes can be introduced into the sequence of an enzyme that can alter the specificity of the catalyzed reaction and/or its catalytic rate. In addition, structural analysis may reveal specific differences in essential enzymes that will enable geneticists to engineer protein sequences, or drug designers to produce reagents, that will selectively counter harmful bacteria or insects without harming the host.

RESEARCH GOALS

The Folding Problem

Inside a cell, amino acids are assembled into peptide chains by a complex system that translates the genetic message into specific amino acid sequences. Following synthesis, the chains fold into compact protein molecules. For many isolated polypeptides, conditions can be provided in which this folding step will occur spontaneously, yielding a biologically active molecule identical to the native, cell-derived protein. Thus, all of the information required to produce the final structure is contained in the amino acid sequence. The prediction of the detailed three-dimensional structure of a protein from a given sequence is known as the folding problem. It is the most fundamental problem at the chemistry-biology interface, and its solution has the highest long-range priority.

The folding problem is not only a major intellectual challenge but also an urgent and immediate problem at the practical level in biotechnology. Successful industrial production of a biologically active protein frequently depends on the ability to induce a cloned polypeptide to fold correctly.

Protein Stability

Protein stability is a specific issue within the folding problem. Thermodynamically, protein structures are only marginally stable, and small changes can substantially increase or decrease their effective stability. Not only is prediction of stability a stringent, but elusive, test of theoretical understanding, but also direct practical applications, through genetic engineering of proteins, are immediately at hand. Resistance to thermal destruction or to degradation by enzymes secreted by microorganisms, for example, are highly desirable properties for pharmacological agents and enzymes in industrial use.

Ligand Binding

The specificity and strength with which ligands, either small or large molecules, bind to proteins is a central feature of biological function. The result of this binding may be simple sequestration for storage or removal, a catalytic event if the protein is an enzyme, the development and transmission of a signal if the protein is a receptor, or the switching off of a gene if it is a repressor. The

ability to predict the structure either of a protein ligand or an active site—the part of the protein where the ligand binds—is central to the rational design of new drugs or new enzymes.

Signal Transmission

Signal transmission from one part of a molecule to another is essential for the regulation of complex enzymes and for the activity of many receptor systems. Its mechanism, however, is poorly understood. Changes in the dynamic properties of an entire protein molecule can be caused by the binding of a ligand to a relatively small active site; this phenomenon has clear implications for information transfer. In some proteins, however, similar interactions produce only localized changes in structure or dynamics. Full understanding will only come through a detailed study of the relevant structures and their properties.

RECENT ADVANCES IN KNOWLEDGE

About 300 protein structures are known, and about 10 to 20 new structures are reported each year. In many cases, knowledge of these structures has brought us closer to the goals outlined above—understanding folding, stability, ligand binding, signal transmission, and catalytic activity. A few examples will illustrate what has been gleaned from studies of structure and the potential applications of the knowledge.

Angiotensin is a chemical in human blood that is involved in regulation of blood pressure. When it is modified by an enzyme known as angiotensin-converting enzyme (ACE), it causes a rapid increase in blood pressure. Control of high blood pressure seemed possible if an inhibitor of ACE could be found. At the time a drug development program was started, ACE had not been isolated in pure form from humans. But the structure of an enzyme from the pancreas of cows, which happens to catalyze a chemically similar reaction, was known at high resolution. Based on the detailed structure of the animal enzyme, especially the catalytic binding site, it was possible to make a model for the active site of human ACE. With this model, the drug captopril, a strong inhibitor of human ACE, was successfully designed and synthesized. Subsequently, protein chemistry was used to design enalapril, a new drug in which some of the unwanted side effects of captopril have been eliminated.

Influenza virus offers another example. This virus causes recurring epidemics (and the continuing need for development of new vaccines) because its surface proteins vary so much. Recently, the structure of haemagglutinin, a surface protein, was determined. Consequently, the regions that vary with the strain of virus have been located. Even more promising is the discovery of a region that does not vary; it provides a pocket in the protein and may be an excellent target for drug development. The full high-resolution structure was essential to the discovery of this region. Recent structural studies of the cold and polio viruses and adenoviruses have opened up similar exciting opportunities.

The recent determination of the structure of a part of an enzyme called DNA polymerase I, in conjunction with related kinetic studies, is beginning to lay a general foundation for understanding how “processive” enzyme reactions work. Such enzymes latch onto a long molecule (the DNA thread in the case of DNA polymerase) and then move rapidly along it without letting go, much like a train on a track. Such a mechanism is entirely new in the field of enzymology. Not only is this fascinating to biochemists studying the replication of DNA and RNA—of central importance to life—but the practical importance is considerable. The processive digestion of polysaccharides and other macromolecules is of great importance to food processing and pharmaceutical industries, for example.

Instrumentation developments have been essential to these and numerous other examples of major progress in structural biology. New or improved instrumentation has led to tremendous savings of time and manpower, as well as to unique routes for solving research problems. Accordingly, much of the remainder of this report will focus on research advances in instrumentation and data analysis.

TECHNOLOGICAL ADVANCES

X-Ray Diffraction

Since the early 1960s, x-ray crystallography has continued to provide us with the most detailed and comprehensive picture of three-dimensional protein structure. When x-rays are passed through a crystal, they are diffracted in many directions, and the geometry and intensity of the many diffracted beams are directly related to the structure of the crystal. Improvements in x-ray sources, in data collection equipment, and in the power and availability of computers are expected to continue to enhance the power of these structure studies.

Area Detectors

The diffraction pattern from a crystal can be recorded all at once (with a photographic film) or one beam at a time (with an appropriate counter). Data on film must be read optically and then must be converted to a digital form to determine the intensity of each beam. Area detectors are a major new innovation in such data collection. They have the advantage of direct counting while retaining the multiple recording capability of film—with a higher signal-to-noise ratio. Analyses that took weeks or months have been reduced to hours, with a considerable gain in accuracy.

This time-saving device has had far-reaching effects on the kinds of experiments that are being planned. For example, it will now be possible to take full advantage of the ability of molecular genetics to generate many different mutational changes in a single protein. The crystallographic examination of the different forms of a given protein will be practical for a small group of investigators, or even a single individual. The comparison of structures will be of inestimable value in elucidating determinants of structure and structure/function relationships.

Synchrotron X-Ray Sources

Synchrotron x-ray sources are becoming available at various national facilities, and their accessibility provides unique opportunities. Because synchrotron radiation has a continuously varying wavelength, it is possible to collect data at two or more different wavelengths. Proper combination of the data sets provides substantial help in overcoming the major stumbling block in solving an unknown structure.

The high intensity of the synchrotron radiation also far exceeds that of any usual laboratory source. It is possible to collect enough data for structure definition in 10 to 100 milliseconds, and perhaps even faster in the future. With such high data rates, full structural studies of short-lived intermediates in enzyme reactions are possible in principle. The determination of the structure of such intermediates at physiological temperatures would lead to a dramatic improvement of our understanding of enzyme catalysis. Currently, intermediate states can only be studied when stabilized under unusual conditions, such as very low temperature, so there is always uncertainty about the relevance of any findings to the actual catalytic process.

The full benefits of fast synchrotron data collection are not yet being realized because the diffracted intensities are recorded on photographic film. More effective use of synchrotron facilities for protein structure investigations will require a *high-flux area de*

tector capable of recording perhaps 10^8 events per second. This is at least 1,000 times faster than the commercial instruments now available for laboratory use.

Computing and Graphics

X-ray studies require computers for data analysis and interpretation. Model refinement to get the "best" structure is particularly computer-intensive. Computer-aided molecular graphics plays an increasing role in the solution of structure and in subsequent study of the structure. Color is routine and is now central to the effective use of graphics as a laboratory tool. It is possible to examine very complex structures and to selectively "flag" special features or properties by color-coding the atoms. Spatial relations that would be extremely difficult to detect by computation become very obvious to the human eye. The emphasis on graphics is likely to continue, even with marked improvements in automatic data analysis, which is itself highly computer-intensive.

Neutron Diffraction

Neutron diffraction has a special role in structure studies because of its unique ability to reveal the position of hydrogen atoms. These atoms are frequently of central importance in enzyme-catalyzed reactions, but because hydrogen is so light, it is poorly detected—if at all—in x-ray structures of proteins. However, hydrogen is easily "seen" by neutrons, although such experiments can only be carried out at the national laboratory reactors. Together with the hoped-for upgrading of the high-flux reactors, the development and capabilities of the new pulse neutron source at Los Alamos will be watched with great interest.

Crystallization

Protein crystals are unusual examples of the solid state in that they contain a large amount of liquid water. The structure of those proteins for which comparisons have been made is essentially identical in a crystal and in aqueous solution.

Production of the highly ordered crystals required for x-ray diffraction studies remains an art rather than a science. Nonetheless, the number of crystallized proteins is increasing rapidly. Of special note is the recent successful crystallization of integral membrane proteins such as the photosynthetic reaction center and bacterial rhodopsin.

Nuclear Magnetic Resonance

With a nuclear magnetic resonance (NMR) spectrum, researchers measure the absorption of radio-frequency energy by the nuclei of molecules placed in a magnetic field. The frequency at which an atomic nucleus absorbs radiation is very sensitive to the chemical environment provided by the structure of the molecule. A basic problem, however, has been the identification of which peak in the spectrum belongs to which atomic nucleus in the structure. New data collection techniques have been developed to extract information about the distances between neighboring atoms, and these techniques have revolutionized the study of proteins up to approximately 12,000 in molecular weight. From these identified spectra, scientists can derive characteristic patterns of substructures in the protein. More detailed analysis with computer-intensive distance geometry algorithms can provide the full three-dimensional structure in favorable cases.

The recent progress in research on small proteins has been directed toward the determination of average structures *in solution* for comparison with models from x-ray diffraction. This work has set the stage for the next fascinating phase of NMR, the study of changes in structure that are induced, for example, by the binding of ligands. While much of this work will be done in partner

ship with investigators using x-ray diffraction, many problems are only accessible through NMR procedures—notably those cases in which crystalline materials cannot be obtained. Unique opportunities exist to learn about partially or totally disordered molecules that are important both in equilibrium populations and as reaction intermediates.

Defining procedures for the precise manipulation of nuclear spin in a molecule—spin engineering—will continue to play an important role in the development of operating procedures for NMR spectrometers, especially for macromolecules with their complex spin systems. An appropriate sequence of radio-frequency pulses can drastically simplify a complex spectrum, reveal relations between spatially distant atoms, and greatly assist in the essential step of assigning peak signals to portions of the protein structure. Further developments require that young investigators with a background in quantum physics be attracted to this particular area of structural biology.

Several other techniques designed for structures larger than those with molecular weights of 15,000 to 20,000 have also been developed. Solid-state NMR has no inherent size limit, and there are very interesting applications for membrane proteins or fibrous materials, such as collagen, which are intrinsically insoluble. Another approach is the direct study of small substrates or inhibitors interacting with active sites of large enzymes. A number of new developments are being intensively pursued in this area, such as the use of labeled, tightly bound substrates. A third approach is to simplify the spectra by preparing samples with stable isotopes inserted in a limited number of known positions. By a combination of chemical and biological procedures, amino acids are prepared with the isotopes ^{2}H, ^{13}C, or ^{15}N in appropriate positions. These are incorporated into proteins at known locations. The syntheses are often difficult, but the rewards are great because the spectra of the isotopically substituted proteins can be very simple and easy to interpret. Moreover, this technique can produce very large signals in comparison with the low background absorption. Even low concentrations of relatively unstable intermediates, such as are likely to be important in the protein folding problem, may be detectable in these enriched samples. And as a further benefit, data collection times are markedly shortened.

Synthesis of Proteins

Sensitivity and sample size continue to limit both x-ray crystallography and NMR, which require amounts of material in the 10- to 100-mg range. An adequate quantity of highly purified proteins of specified sequence, and, where required, with specific isotopic substitutions, is essential to biophysical study of structure and function. Within different but overlapping size ranges, quantities of proteins can be produced today either by chemical or biological procedures.

Chemical Synthesis

Solid-phase chemical synthesis has been effectively automated, and peptides 30 to 40 amino acids long are readily produced in good yield. Substantially longer peptides also have been synthesized, and continuing improvement can be expected. The next step toward the synthesis of longer chains is the condensation of preformed fragments. Condensation is possible by enzymatic as well as chemical methods, but general procedures are not as well worked out and deserve considerably more study.

Chemical synthesis allows the insertion of an isotopically labeled amino acid, an amino acid derivative, or even a nonnatural amino acid in any single position in the chain. Multiple-site “mutations” at any selected group of sites thus become easy to produce. In the synthesis of drugs that mimic pep

tides, even the peptide bond may be circumvented in specified locations, leading to resistance to degradation.

Limitations in chemical synthesis at this time derive from the problem of optical purity, the yield of the correct sequence, and the roughly linear relationship between the amount of the product and the cost of producing it. However, these limitations, as well as the limitation on chain length, are overcome in some biotechnology processes.

Biological Synthesis

The procedures that are used to produce proteins in high yield by cloning in bacteria are well developed. The companion procedures for producing single amino acid changes by site-directed mutagenesis are simple, fast, and reliable. In many cases, the fusion of a special "signal sequence" to the protein will result in its secretion, which assists in proper folding. Nonetheless, cloning is not always successful. Degradation of the product can severely reduce the yield; inside the bacterial cell, the chain may not fold to yield the desired, biologically active molecule; and modification of certain amino acids after polymerization of the protein as required for various eukaryotic proteins ("post-translational" modification) does not occur in bacteria.

Cloning in eukaryotic systems, particularly human cell culture, is not yet as well developed and is very expensive for producing proteins in commercial quantities. Improved expression vectors to overproduce the desired protein, protease-deficient strains, rapid lysis methods, and better biochemical separation procedures are all required. *Two important research opportunities are development of better methods for multiple site-directed changes at positions widely separated in the sequence, and the development of high expression systems for eukaryotic proteins.*

It is equally important that these new protein products be characterized rapidly. Preliminary structure evaluations by standard biophysical procedures, particularly the various forms of optical spectroscopy, enable screening of protein products and selection of the most interesting for detailed study by x-ray and NMR techniques.

Theory

Theoretical studies of proteins are only beginning to have a real impact in relation to biological function. The field has been stimulated by advances in computer technology, theoretical chemistry, and experimental biochemistry. It is clear that theoretical studies will play a major role in the design of new proteins and of molecules with which proteins interact.

The most highly developed theoretical methods involve molecular dynamics simulations, in which computers are used to simulate atomic motions in a protein and its surroundings. When combined with a new approach giving thermodynamic parameters for reactions, dynamic simulations can be used to make predictions concerning recognition and binding among proteins and other molecules. This method has recently been used successfully to calculate the affinity in an enzyme-inhibitor interaction; it has promise for studies of protein folding, stability, covalent reactivity, and noncovalent association. Practical applications include the design of drugs, enzymes, antibodies, and other molecules.

The rates and mechanisms of enzyme-catalyzed reactions and ligand binding are potentially accessible through molecular dynamics. A modification known as Brownian dynamics is useful for extending calculations into the time range of somewhat slower biological processes. A number of other theoretical approaches, clearly on the horizon, may be useful for predicting the structures of short peptides in solution.

Continuing attention should be directed to improving the basic mathematical functions and input parameters that are needed both in molecular dynamics and energy

minimization procedures; to improving the treatment of electrostatic interactions; and to the detailed treatment of the water—protein interface where biological activity is expressed. More ad hoc approaches in applying other aspects of basic chemical theory may also be useful in attacking the protein folding problem in which the proper application of first principles is still elusive.

BOTTLENECKS AND RECOMMENDATIONS

Progress in solving the scientific problems of structural biology will require both personnel qualified in this multidisciplinary area and sophisticated equipment in individual laboratories and national centers. For structural biology to achieve its full potential in contributing to fundamental science, medicine, agriculture, and the chemical industries, a number of policy concerns should be addressed.

1. *Basic Research Support* Of absolutely central concern in this area, as in others, is *maintenance of support for basic research programs* at the level of the individual investigator and small consortia.
2. *Professional Personnel* Of comparable concern is the continuing *supply and training of professional personnel.* Scientific opportunities will not be realized, and the equipment initiatives suggested below will have little effect, if trained personnel are not available. In the recent past, there has been a relatively small number of individuals entering biophysics, biophysical chemistry, and the general field of structural biology. The major attraction during that period was clearly molecular genetics. During the past year or two, there has been an increasing number of entering graduate students interested in quantitative structural studies.

   This student interest has coincided with a dramatic upsurge in activity in the industrial sector. A substantial number of biotechnology firms have set up structural units, including x-ray crystallography, high-resolution NMR, and theoretical modeling. Researchers experienced in one or more aspects of structural biology and general protein chemistry are in high demand at this time, and corporations have attracted much of the presently available talent.

   Currently available predoctoral training programs may be adequate to provide the necessary graduate student input to this field *provided these predoctoral training programs are maintained at a level at least equal to their present levels.* A particularly effective source of highly qualified personnel is provided by the Medical Scientist Training Program. If there are further cuts in any of the programs, the structure area, which is poised scientifically for substantial progress, will be nipped in the bud, and will suffer proportionally more than the well-populated areas of molecular and cellular biology.

   The most serious concern for personnel is postdoctoral training. The new generation of structural biologists must be familiar with molecular biology as well as biophysics, and it would be highly desirable that molecular biologists with an interest in structure learn at least the rudiments of the biophysical methods. Similarly, physicists and instrumentation engineers whose expertise could be shifted rapidly to structural biology must learn some molecular biology. This interdisciplinary training is difficult to accomplish properly in the time period of a normal doctoral program; postdoctoral training thus becomes even more essential than usual under these circumstances. *We urge that the postdoctoral fellowship programs be maintained and, if possible, expanded to cover the present and anticipated needs in structural biology.*
3. *Supply and Development of Major Instruments* The entire field of structural biology is heavily dependent on major equipment items and on unique facilities available at certain national centers. Continuing progress on the biological problems in this field will be closely correlated with the

improvement and availability of advanced instrumentation.

X-Ray Diffraction Currently available area detectors are having an enormous impact on the efficiency of data collection and on the types of research programs that it is realistic to plan. *Agencies should be prepared to fund acquisition of area detectors, and ancillary equipment for efficient utilization, widely throughout the structural biology community over the next few years.*

The effective use of the synchrotron x-ray sources at the national laboratories will depend on the development of high-flux area detectors capable of recording at least 10^8 events per second. Efforts are under way for other scientific fields; the needs of structural biology should be considered as part of this general development effort.

Nuclear Magnetic Resonance The distribution of 500-MHz instruments, or their future replacements, will continue to present a policy problem. Laboratories devoted to the development of NMR techniques, or to major long-term protein-structure projects, will need fully dedicated instruments of their own. Shared facilities should still be dedicated to the study of macromolecules and should not be expected to provide small molecule spectra as an additional service component.

Improvements in resolution and sensitivity will depend on the development of stronger magnets. Sensitivity and resolution both become increasingly important as the size of the protein increases, and both are improved as the magnetic field strength of the spectrometer is increased. *We suggest that a major effort be launched to interest and encourage instrument companies to produce spectrometers with a 17.5 T (750-MHz) magnet.* This appears to be feasible with currently available superconducting wire technology, although some engineering difficulties remain to be solved. A magnet at 20–25 T (~1,000 MHz) is not out of the question, although an intensive investigation in the materials science area may be involved. Success, however, would have a dramatic impact on biophysical NMR studies.

Availability of precursor chemicals labeled with stable isotopes is another NMR concern. The very high price of labeled material is a serious general problem for scientists. The Stable Isotope Facility at Los Alamos is bound by regulation to stop any activity that is taken up in the private sector. In contrast to many other examples of scientific resource supply, this particular regulation has worked to the disadvantage of the research community. There is not a sufficiently large market for labeled compounds to reduce the price through volume and competition. Even the development of clinical applications would produce a market for only a relatively small number of compounds and would not cover the broad range needed for the research proposed above. *It is essential that a mechanism be found for providing amino acids and nucleotides with a variety of different stable isotope labeling patterns for the research community.*

Computers The instrumentation needs for theoretical and experimental studies of proteins include *increased and predictable access to supercomputers;* improved communications between these machines and remote users through networking; and additional access to high-quality graphics devices, scientific workstations, and special-purpose computers.

X-ray crystallography, NMR, and theoretical studies all have very heavy computing requirements. The larger the molecules and the faster the data collection, the larger will be the computational requirements in the immediate future. Even for structures of modest size, the current iterative x-ray refinement procedures create major inconvenience to other users of a VAX. (Refinement cycles can occupy tens of hours per cycle.) Protein data processing even from current NMR spectrometers requires levels of computation not normally available in the laboratories of individual investigators. Many theoretical

problems are not at all practical on intermediate-level machines.

We estimate that the needs of the current scientific community in structural biology nationwide may already exceed the computing power represented by two advanced-level Cray machines. The present National Science Foundation Supercomputer Center Program is useful in providing access to these machines. This initiative, however, covers all areas of science and may well become saturated. The necessary phased expansion of this program should include the present and anticipated requirements of the structural biologists. A library of programs specifically written to take advantage of the architecture of these machines also will be essential for their effective use.

The efficient use of the supercomputers will depend on the ease and convenience of access. The latter will depend on the speed and effectiveness of the networking that is available to make this possible. Networks suitable for connection of the lower-level computers in the various structure laboratories also will become increasingly useful. Experience over the next year or two with the currently developing general purpose networks will show whether or not they are adequate for this purpose.

A large volume of protein sequence and structure data can be expected in the near future as a result of the many new methodologies. Therefore, it is time to plan for the accession and use of the data in a computerized central data bank. Computer searches of genetic structure data banks have provided significant new insights into biological phenomena, and a similar outcome can be expected from the protein structure data.

Finally, there will be increased need for dedicated microcomputers and various levels of graphics workstations in individual laboratories. Although these are not expected to be particularly expensive, the need for their wide distribution is clearly visible now and should have a prominent place in the planning for future equipment support.

Report of the Research Briefing Panel on Prevention and Treatment of Viral Diseases

Research Briefing Panel on Prevention and Treatment of Viral Diseases

Wolfgang K. Joklik *(Chairman),* Duke University Medical Center, Durham, N.C.
Seymour S. Cohen, State University of New York at Stony Brook (retired), Woods Hole, Mass.
William Haseltine, Dana Farber Cancer Institute, Boston, Mass.
Maurice R. Hilleman, Merck Institute for Therapeutic Research, West Point, Pa.
Joseph L. Melnick, Baylor College of Medicine, Houston, Tex.
Thomas C. Merigan, Jr., Stanford University School of Medicine, Stanford, Calif.
Roland K. Robins, ICN Pharmaceuticals, Inc., Costa Mesa, Calif.
Bernard Roizman, University of Chicago, Chicago, Ill.
Julius S. Youngner, University of Pittsburgh School of Medicine, Pittsburgh, Pa.

Staff

Roy Widdus, *Director,* Division of International Health, Institute of Medicine
Mark Feinberg, *Consultant*
Allan R. Hoffman, *Executive Director,* Committee on Science, Engineering, and Public Policy

Report of the Research Briefing Panel on Prevention and Treatment of Viral Diseases

INTRODUCTION

Viruses are segments of genetic material, either RNA or DNA, encased in protein shells and often further wrapped in lipid-containing envelopes. Viruses multiply only within living cells, commandeering the host cell to synthesize their proteins and also their nucleic acids.

Infection profoundly affects the host cell, bestowing on it characteristics that distinguish it from uninfected cells. Viruses disrupt or kill infected cells or transform them into tumor cells, thereby causing disease. Viral diseases vary in severity from mild and transitory infections to illnesses that terminate in death. Persistent viral infection, which for years may not be accompanied by symptoms, can eventually cause chronic degenerative disease with fatal outcome. Viruses cause a wide variety of cancers in animals, and the epidemiologic and laboratory evidence is very strong that viruses also cause human cancer.

Minimizing the harmful effects of viral infections has long been a major goal of medical and veterinary science. The two major approaches to achieving this goal are preventing the onset of viral diseases by immunization (vaccination) and treating viral diseases by arresting and curing infections once they have started.

Prevention of Viral Diseases

Prevention of viral diseases is based on the fact that viruses usually elicit the formation of protective or “neutralizing” antibodies, cell-mediated immunity, or both. It is therefore possible to protect against viral infection by immunization—raising a host's immune defenses ahead of time—so that when the disease-causing virus enters the host, it is quickly neutralized by antibodies. In conjunction with other components of the host's immune system, these antibodies inactivate, destroy, and eliminate the virus. The immunizing agent may be active virus in the form of a harmless (avirulent) variant, inactivated (killed) virus, or the viral proteins that elicit the formation of neutralizing antibody (subunit vaccine). The earliest form of immunization undoubtedly was variolation, invented many centuries ago by the Chinese, in which persons were exposed to skin scabs from others who had survived smallpox infections. The rationale was that in such cases the disease had been caused

by a less virulent form of the variola or smallpox virus. The practice was dangerous, with a fatality rate of up to 1 percent, but it did afford a measure of protection. The first example of a "killed" vaccine was the rabies vaccine developed by Pasteur. As for the subunit approach, many such vaccines have been devised, but none was widely used in human beings until a hepatitis B virus subunit vaccine was licensed in the United States in 1981.

Prevention of viral diseases has had some outstanding successes. Smallpox, one of the most devastating of all infectious agents in human beings, has been eradicated globally through the use of energetically executed vaccination programs. Yellow fever virus has also essentially been eliminated in some parts of the world. Effective control has been established over poliomyelitis, measles, mumps, and rubella, which until the middle of the twentieth century infected millions of children annually in this country, causing many deaths and disabling even larger numbers.

Treatment of Viral Diseases

Successful treatment of viral diseases requires interruption of virus multiplication by specifically inhibiting the functioning of virus-encoded proteins and nucleic acids. This strategy primarily entails the identification of analogues of nucleic acid and protein components capable of inhibiting virus-specified reactions to a greater extent than reactions essential for host cell multiplication and survival.

The approach used so far has been somewhat empirical: effective compounds are either identified by the use of screening programs, or progressively more active analogues are devised in organic synthesis programs. Numerous drugs have been found that are capable of inhibiting the multiplication of a wide variety of viruses in cultured cells to a greater extent than they inhibit the multiplication of the host cells themselves. Several such drugs (idoxurdine, ribavirin, and Acyclovir) have been licensed for human use.

Empirical approaches are not generally designed to yield drugs that are specific inhibitors of the functions of viral proteins and nucleic acids. Until very recently, most of the drugs examined have been general inhibitors of protein and nucleic acid synthesis, and a few happen to inhibit certain virus-encoded enzymes somewhat more effectively than analogous host cell functions. However, investigation of viral multiplication cycles at the molecular level has greatly expanded knowledge of the virus-specific processes and structures that could serve as targets for antiviral chemotherapy.

Target Viral Diseases

Many viral diseases still present grave problems. The virus that causes acquired immune deficiency syndrome (AIDS) is a recently recognized and serious problem, but many other viruses have long been known to cause a wide spectrum of diseases, both acute and chronic, involving all organ systems of the human body. Diarrheal and respiratory disease viruses are probably the major global cause of morbidity and mortality in children, especially in developing countries. These include rotavirus, parainfluenza viruses, coronaviruses, and respiratory syncytial virus. Influenza virus afflicts all age groups but is most life threatening in the elderly. Other important viruses include hepatitis B, cytomegalovirus, and herpes simplex. Viruses causing hemorrhagic fevers or encephalitis are often insectborne and include dengue virus (prevalent in the Caribbean and Southeast Asia) and Japanese B encephalitis virus (prevalent in the Far East). Poliomyelitis, measles, mumps, and rubella are still widespread in developing countries.

Viruses are suspected of involvement in several slow degenerative diseases; examples include Creutzfeldt-Jakob disease, Alz

heimer's disease, and juvenile diabetes mellitus. The same is true for a variety of human malignancies including cervical carcinoma (certain human papillomaviruses); Burkitt's lymphoma and nasopharyngeal carcinoma (Epstein-Barr virus); liver cancer (hepatitis B virus); and certain human leukemias (HTLV-I and-II). Many of these human tumor viruses were discovered in the past 10 years, and it is expected that more will be found before long.

In addition to the viruses that affect human beings, there are many others that cause disease in animals and plants, with enormous economic loss to agriculture.

In summary, viruses remain among the major scourges of mankind. They cause an enormous burden of illness with resultant economic loss; they kill and permanently disable millions annually both in this country and throughout the world. The objective of this report is to illustrate that recent advances in virology and molecular biology have pointed the way to new strategies for preventing and treating viral diseases.

RECENT ADVANCES IN MOLECULAR VIROLOGY

The advent of two technologies during the past 15 years has led to greater knowledge of the nature of viruses and of their multiplication cycles. These are recombinant DNA (gene-splicing) technology, which permits the isolation and detailed molecular characterization of DNA or RNA, and the technology for producing monoclonal antibodies, which provides reagents not only for specific proteins but also for specific antigenic sites on specific proteins. Application of these methods has provided new insights into the structure of virus particles and surface proteins and into the regulation of virus multiplication. Each of these areas offers different opportunities for combating viral diseases. The more detailed description, below, of advances in each area is followed for each by identification of the specific research opportunities that could be pursued.

THE STRUCTURE OF VIRUS PARTICLES

Great progress has been made in recent years with respect to knowledge of virus structure. The three-dimensional structures of several plant and animal viruses, including some human pathogens (e.g., poliomyelitis virus and rhinovirus) have been determined. These studies are extremely important in the elucidation of how virus particles interact with host cells and antibodies.

These studies on whole virus particles have been paralleled by studies of individual components of virus particle surfaces. Two types of such components are of fundamental importance. The first includes the viral cell attachment proteins, which are the proteins that recognize receptors on susceptible cells. For instance, certain viruses have proteins (termed hemagglutinins) that recognize receptors on erythrocytes (red blood cells) and cause their agglutination. Attachment proteins have been identified for a variety of viruses, and more recently cellular receptors also are being identified. An important question is whether these cellular receptors have some other essential cellular function, in addition to the recognition of viral attachment proteins. Studies of this question and of the role of these receptors during virus contact by and penetration of host cells are under way. They will provide a basis for investigations to determine if the interaction of viruses with their receptors could be prevented, inhibited, or terminated so as to protect from infection.

The second class of viral surface features important to the interaction of viruses and organisms is their antigenic sites, termed epitopes. Here the primary question is which epitopes elicit the formation of protective or neutralizing antibodies. The surface components of many types of virus particles have been elucidated and the genes that encode

many of these proteins have been sequenced, that is, deciphered for the information in their basic structural units. In both cases, studies have included viruses from all the major classes in these two broad categories. The most intensively studied of these proteins is the hemagglutinin (HA) of influenza virus. Not only have the amino acid sequence and three-dimensional structure of the HA of an important influenza strain been determined, but variant strains (with different antigenic properties) also have been sequenced. As a result, it is now known where the important epitopes on the influenza virus HA are, what the nature of the antibodies synthesized in response to individual epitopes is, and how single amino acid changes affect epitope function. Similar studies are in progress on several other viruses and viral surface components.

SPECIFIC STRATEGIES FOR PREVENTING VIRAL DISEASES

Strategies for preventing viral diseases have as their goal the neutralization of virus particles before they can establish productive infection. Among the most interesting approaches are those discussed below.

Genetically Modified Live Viruses.

The most successful vaccines are those containing attenuated (weakened) live virus particles. Attenuation is generally achieved by serially culturing the virulent virus in the cells or tissues of some other host, which often gives rise to variants that have lost virulence for human beings. Suitable variant strains that still will grow well enough in the human host to elicit the formation of neutralizing antibodies against the disease-causing strain, but that cause neither disease nor untoward reactions, are then selected. It has recently become clear that genetic engineering techniques could be used to provide improved attenuated vaccine strains much more rapidly than the slow and relatively uncontrollable process of serial culturing.

The genetic material of many viruses has now been sequenced, and the genes responsible for specifying the interactions of the virus with the host organism are being mapped. These include the genes responsible for determining affinity for particular host tissues, ability of the virus to spread from one location in the organism to another, virulence and nature of cytopathic effects, capacity to establish latent or persistent infection, and immunogenicity. A wide variety of genetic techniques are available to identify and manipulate such genes. Once the genes governing virulence or other factors have been identified and characterized, they can be altered or inactivated relatively simply. In this manner, it should be possible to provide a new generation of acceptably safe vaccine virus strains.

The Use of Virus Vectors

The genes for many proteins capable of eliciting the formation of neutralizing antibodies have been isolated. Such genes derived from virulent disease-causing virus can be inserted into avirulent vectors such as vaccinia virus or adenovirus. When these vectors are used to infect hosts (without causing disease), the inserted foreign genes are expressed, and the host develops antibodies and immunity to the virus from which they were derived. The feasibility of this approach has been demonstrated: a vector carrying the major rabies virus glycoprotein has been used successfully to protect foxes against challenge with wild virus. Work is needed to optimize the safety and efficiency of the vectors.

Purified Viral Proteins as Antigens for Vaccines

When individual proteins that elicit the formation of neutralizing antibodies were

first identified, attempts were begun to use them as subunit vaccines. Until recently the major difficulty was the inability to isolate sufficiently large amounts of the proteins in a state of sufficient purity. Molecular cloning techniques have greatly improved the feasibility of this approach. Genes for viral proteins (such as those for protective antigens) can now be inserted into a variety of prokaryotic and eukaryotic expression vectors. These vectors can be introduced into bacterial, yeast, or mammalian cells where they can be made to induce the synthesis of large amounts of the specific proteins according to instructions coded by the inserted viral gene. These viral proteins could then be purified in large quantities and used as safe and specific vaccines.

Development of Technologies for Enhancing Immunogenicity

An important aspect of antibody formation is the optimal presentation of antigens to the immune system. Recent advances in immunology have suggested several new ways of improving such presentation. Among possible approaches are the use of protein conjugates or aggregates, liposomes, or immunostimulatory complexes produced with plant extracts.

STRATEGIES FOR THE TREATMENT OF VIRAL DISEASES

Strategies for the treatment of viral diseases have as their aim the interruption of viral infections once they have started, resulting in the elimination of the virus from the body and a cure of the disease. This aim can be addressed best with detailed knowledge of the key reactions of viral multiplication cycles. Such reactions are catalyzed by virus-specified enzymes engaged in subtle interactions of viral and cellular proteins and nucleic acids. In the case of enveloped viruses, interactions of viral and cellular proteins with lipid membranes are also of crucial importance. In recent years, technical advances provided by recombinant DNA technology and the availability of monoclonal antibodies have led to a dramatic increase in knowledge of the processes involved in virus multiplication and the interaction of viruses with their host cells. The following is an outline of current knowledge about virus multiplication and of the opportunities to inhibit it.

The Viral Multiplication Cycle

Viruses multiply by means of precisely regulated series of reactions. Typically, viruses adhere to host cells via specific receptors and are internalized by a process of engulfment (phagocytosis). The viral genome (its DNA or RNA) is then liberated from its protective protein coat, and the viral genetic information is expressed through messenger RNAs that are translated into proteins. Viruses vary in their complexity; thus the number of proteins encoded by viruses varies: some encode fewer than 10, others more than 100. Subsequently the viral nucleic acid replicates, and the newly replicated (progeny) viral genomes are enclosed within newly formed protein coats. The number of progeny virus particles formed in a cell may vary from a few hundred to more than 100,000. Progeny virus particles are liberated either when the host cell disintegrates or when virus particles “bud” through the cell membrane, thereby acquiring their envelopes.

Strategies for the Expression of Viral Genetic Information

One of the key steps in viral multiplication is the expression of the genetic information encoded in the viral genome. To achieve this, different viruses use different strategies. The genetic information of single-stranded RNA viruses is translated into proteins either directly or through the involvement of a second messenger strand of

RNA complementary to the genome. The latter process involves a virus-encoded enzyme present in the virus particle. RNA viruses containing double-stranded RNA must also first be transcribed into messenger RNA by virus-encoded enzymes present in virus particles.

RNA viruses also include the retroviruses. Upon infection, their RNA is transcribed by a virus-encoded RNA-dependent DNA polymerase (reverse transcriptase) into double-stranded DNA (the provirus). Another unique enzyme newly translated from the viral RNA integrates this provirus into the host cell nucleic acid. The integrated provirus directs host cell enzymes in the synthesis of viral proteins and viral RNA, from which new progeny viruses are assembled.

DNA-containing viruses can only express their genetic information by its being transcribed into messenger RNA. Some use host enzymes for this purpose; others specify their own DNA-dependent RNA polymerases.

Strategies for the Replication of Viral Genetic Material

Viruses employ various strategies for replicating their genetic material just as they do various strategies for expressing it. The replication of all RNA viruses except retroviruses involves virus-encoded enzymes because uninfected cells do not possess enzymes capable of replicating RNA. The replication of DNA genomes is accomplished either by host cell DNA polymerases, or by virus-encoded DNA polymerases.

Opportunities for Interfering with the Viral Growth Cycle

Recent advances in molecular virology have led to the following picture of viruses. Their genomes comprise both regulatory regions and coding regions. The regulatory regions include sequences that serve to regulate a variety of processes, such as nucleic acid and protein synthesis and virus assembly, in a variety of ways (e.g., recognition, initiation, promotion, enhancement, and termination). These regulatory regions of viral nucleic acids may function by interacting with proteins or by interacting with other nucleic acid sequences. The coding regions encode virus-specified proteins. Viral proteins are of three kinds: (1) structural components of virus particles, (2) enzymes, and (3) regulatory proteins that interact with nucleic acids or other proteins.

Numerous opportunities exist for inhibiting virus multiplication. One approach is to inhibit the activity of some viral enzymes; another is to disrupt the action of some regulatory protein; and a third is to interfere with the function of a regulatory nucleic acid sequence. Such approaches are made possible by the fact that the nucleic acid of many viruses has now been molecularly cloned and sequenced. As a result, not only are the sequences of many viral proteins now known, but also the sequences of many regions of nucleic acid with regulatory functions.

The following appear to be feasible strategies for inhibiting virus multiplication.

Inhibition of Virus-Encoded Enzymes

Nucleic Acid Synthesis Inhibition of viral nucleic acid synthesis would interrupt viral multiplication and infection. The feasibility of this approach is confirmed by the fact that most successful antiviral compounds currently licensed (e.g., Acyclovir) or under investigation are nucleotide analogues. Because host and viral enzymes differ in their ability to use these compounds as substrates, they (or their metabolites) are preferentially incorporated into viral nucleic acids and halt multiplication. Virus-encoded enzymes of nucleic acid synthesis—the RNA and DNA polymerases—are, therefore, the most obvious targets for antiviral chemotherapy. Many, like the RNA-dependent RNA polymerases, have no counterparts in uninfected cells. Now that genes for many of the viral enzymes of nucleic acid synthesis have

been cloned, it will soon be possible to prepare them in large amounts. When the structures of the enzymes' catalytic sites have been determined, it should be possible to design compounds, possibly nucleotide analogues, that irreversibly inhibit them specifically.

Proteases These enzymes catalyze reactions essential to viral multiplication. Many viral proteins are synthesized in the form of precursors, usually 20 to 50 percent larger than the functional proteins. Sometimes several proteins are synthesized linked together in the form of a polyprotein. Precursors and polyproteins are cleaved by highly specific virus-encoded proteases to active individual proteins. Proteases are, therefore, potential targets for antiviral chemotherapy.

Capping Enzymes Several viruses encode enzymes that form caps (modified single nucleotide additions) at the end of messenger RNA molecules that are essential to the efficient functioning of these molecules. The viral cap-synthesizing enzymes are analogous to corresponding host cell enzymes, but their amino acid sequences are likely to be quite different. Therefore, they represent a unique target for intervention.

Integrases The genomes of several types of virus are inserted into host cell DNA as an essential part of the virus life cycle. Usually, but not invariably, this is the first step of transforming normal cells into tumor cells. Some viruses use host cell enzymes for this purpose, but others, such as the retroviruses, encode their own integrase enzyme for this purpose. These enzymes apparently are not capable of recognizing unique cellular DNA sequences; but often the nucleic acid segments that are integrated, such as retroviral proviruses, possess highly distinctive features recognized by integrases. This recognition feature of integrases presents a target for antiviral chemotherapy.

Sequence-Specific Nucleases Virus-encoded, sequence-specific nucleases perform essential functions during viral genome replication; that is, they cut nucleic acids at precisely specified positions. Again, these represent a selective target.

Inhibition of Interactions of Viral and Host Cell Proteins with Regulatory Sequences in Viral Genomes

Precise regulation is essential to the complex process of viral multiplication. Most, if not all, regulatory regions in viral genomes operate through interaction with proteins that have the ability to recognize and interact with specific nucleic acid sequences (i.e., they are sequence-specific binding proteins). Such highly specific nucleic acid-protein interactions promise to be targets for antiviral chemotherapy, but present rudimentary knowledge of these interactions make this a long-term approach. Two strategies can be imagined. First, it is now becoming feasible to identify the regions of proteins that bind to nucleic acids. From such proteins that bind to nucleic acids. From such proteins, it may be possible to isolate peptides that retain nucleic acid binding ability and to use either these peptides, or analogues that bind even more strongly, to saturate the regulatory sequences, thus rendering them unavailable to the functional viral proteins. The second approach would be to use short complementary nucleic acid sequences to saturate the regulatory sequences.

Interference with Messenger RNA Function

Translation of viral genetic information, by means of messenger RNA (mRNA), into proteins is essential to viral multiplication, and for this the mRNA must be accessible to the protein-synthesizing machinery as a single-stranded RNA. Through genetic engineering techniques, it is possible to synthesize RNA segments that are complementary to mRNA and that bind to

it, blocking translation. The present question regarding this approach is how to introduce the inhibitory RNA into cells. However, new ideas are constantly being conceptualized and evaluated experimentally. Thus, this may be a feasible long-term approach to antiviral chemotherapy.

The Target-Cell Approach

A major concern of antiviral chemotherapy is that even in the most severe diseases only a very small fraction of cells in an organism are infected. Clearly it would be advantageous to aim the antiviral agent at the infected cell rather than to introduce it into every cell. Attention has therefore been directed at various forms of a target-cell approach to antiviral chemotherapy. At least three strategies are being explored. The first and most attractive is exemplified by a strategy that is feasible in cells infected with some, but not all, herpesviruses. These viruses encode an enzyme, a deoxypyrimidine kinase, that phosphorylates nucleoside analogues (such as Acyclovir) not readily phosphorylated by the analogous host cell enzyme (thymidine kinase). Such phosphorylated nucleoside analogues are then incorporated into viral DNA and inhibit virus multiplication. (Phosphorylation of nucleosides is a prerequisite for incorporation into viral DNA.) The attractive feature of this approach is that the antiviral nucleoside analogues are not phosphorylated in uninfected cells and therefore only exert their inhibitory effects in infected cells.

The second target-cell approach takes advantage of the fact that virus-encoded proteins are incorporated into the membranes of infected host cells soon after infection. Monoclonal antibodies against such proteins can be produced and coupled to inhibitors of nucleic acid or protein synthesis that would normally not enter cells. When combined with antibody molecules, however, they are internalized. Thus, nonspecific inhibitors of nucleic acid and protein synthesis can be introduced specifically into infected cells.

The third target-cell approach is predicated on the fact that soon after infection the permeability of the cell membrane often increases, potentially permitting the entry of compounds that would be excluded from uninfected cells.

Several other strategies for inhibiting viral infections can be envisaged. Possible strategies include direct inhibition of virus-encoded regulatory proteins, inhibition of the interaction of viruses with their cellular receptors, inhibition of the budding process for the release of enveloped viruses from infected cells, inhibition of viral protein glycolysation (addition of sugar residues), and inhibition of the intracellular transport of viral proteins. The research necessary to determine the feasibility of these strategies can now be outlined in some detail.

THE NEED FOR IMPROVED TECHNIQUES TO DIAGNOSE VIRAL INFECTIONS

There is a pressing need for ways to rapidly diagnose viral infections. Treatment with a specific antiviral agent cannot be selected before the infecting virus is identified. Recent advances in biotechnology such as nucleic acid probe techniques and monoclonal antibodies have enhanced capabilities in this area, and excellent progress has been made on some problems, such as determining exposure to the virus that causes AIDS. The likely availability of effective antiviral drugs should provide a stimulus to the development of diagnostic tools.

SELECTED OPPORTUNITIES AMONG VIRAL DISEASES

The following viral diseases are high-priority areas for research into their prevention and treatment.

Acquired Immune Deficiency Syndrome (AIDS)

The severity of the AIDS problem warrants major efforts in prevention and treatment, but immediate prospects for either are not highly promising. In this circumstance a number of approaches should be pursued, and research on pathogenesis of the disease should be actively continued. It is not yet clear how best to approach the development of a vaccine to prevent AIDS. Among questions to be answered are whether antibodies to the major viral surface protein can be protective, the significance of the genetic variability of the virus, and why natural infection does not elicit protective antibodies.

Control of the persistent infection occurring with the AIDS virus is also problematic; drugs designed against it might have to be taken for extended periods. The virus encodes three enzymes, known for a number of years to exist in other retroviruses but not yet characterized: a reverse transcriptase, an integrase, and a protease. All of these are obvious targets for rational drug design efforts along with two newly identified regulatory proteins. Once the AIDS provirus has been integrated into the host cell DNA, drugs directed against any of these targets might have to be taken for the lifetime of the patient to prevent the spread of the virus, if the virus were not eliminated from the body by the drug. It is likely that this approach would only be practical and feasible if a strategy of targeting infected cells were adopted. Measures to remove the AIDS provirus from the genome of infected cells cannot yet be formulated.

Further specific recommendations will be made in the report of the National Academy of Sciences-Institute of Medicine Committee on a National Strategy for AIDS, planned for publication in September 1986.

Influenza Viruses

The optimal strategy for protection against influenza viruses would be a safe and effective vaccine, and new approaches are constantly being tested. In addition, drugs against influenza virus would be desirable when new virus strains pathogenic for humans appear, before adequate amounts of new vaccine can be prepared. Targets for anti-influenza virus chemotherapy are an RNA-dependent RNA polymerase and a unique “cap stealing” protein involved in nucleic acid synthesis. Similar considerations apply to disease caused by respiratory syncytial virus and the parainfluenzaviruses, which are also important human pathogens.

Herpes Simplex Viruses (HSV) 1 and 2

A few groups are pursuing various (subunit and genetically engineered attenuated) approaches to development of a vaccine to prevent this disease. However, because latent infections can give rise to recurrence of symptoms even in the presence of antibody, a vaccine may not prevent infection but rather reduce the severity of initial lesions, their recurrences, and possibly their frequency.

Intensive, long-term research will be necessary to devise ways to eliminate latent virus once infection has occurred. However, herpes simplex virus (HSV) presents numerous targets for antiviral drugs to treat the lesions and other symptoms that occur with the initial infection and its recurrences. The best is probably deoxypyrimidine kinase, which invites the target-cell approach described above; progressively more effective nucleoside analogues (that function like Acyclovir) are constantly being synthesized. Other good targets are provided by the DNA polymerase and ribonucleotide reductase that are encoded by these viruses.

HSV 1 and 2 can serve as models for controlling human infections with other herpesviruses (cytomegalovirus, Epstein-Barr

virus, and varicella-zoster virus). However, each of these presents unique epidemiologic, economic, and disease problems. Probably the most costly in terms of percentage of total health costs are cytomegalovirus infections in immunocompromised people, such as transplant recipients, and in pregnant women, leading to mental and developmental retardation of their off-spring. Epstein-Barr virus in the United States is associated with disseminated mild to severe infections of young adults (mononucleosis), severe infections in immunocompromised individuals, and fulminating lethal infections in a small number of children. Outside the United States it is associated with certain malignancies including nasopharyngeal cancer. Varicella-zoster virus is the agent of chickenpox in children and shingles in adults.

For all these herpesviruses, specific strategies can be devised for vaccines or antivirals; for example, a vaccine for varicella-zoster should be genetically engineered so that the genes that enable it to initiate latent infections (that may result in shingles) are removed.

Hepatitis B Virus (HBV)

Worldwide, millions of people are carriers of hepatitis B virus (HBV); that is, they are persistently and chronically infected with this virus. Each year 800,000 persons, mostly in developing countries, die of the consequences of HBV infection (cirrhosis and liver cancer). For these persons, antivirals would be helpful, and HBV encodes a unique DNA polymerase that would be an excellent target for chemotherapy. For those not yet infected, HBV is highly amenable to new vaccine development approaches; a plasma-derived vaccine is available but expensive. The surface antigen of HBV has been cloned and can be expressed in yeast cells. In highly purified form, it is being tested for its ability to elicit the formation of neutralizing antibodies. Another promising approach to the prevention of HBV infection is the insertion of the gene for the HBV surface antigen into a vector virus such as vaccinia virus. Immunization of newborns is the favored prevention strategy in areas of high incidence of disease.

Rotaviruses

Rotaviruses are an important cause of diarrheal disease and infant mortality worldwide. Both the antiviral drug approach and the vaccine approach appear to be promising. The rotavirus multiplication cycle involves two different RNA-dependent RNA polymerases, which are therefore good targets for antiviral chemotherapy.

Some rotavirus vaccine candidates are in development, such as those based on bovine or rhesus rotaviruses; other candidates have been developed through genetic reassortment techniques. However, the genes that encode rotavirus surface antigens have been cloned and are now being placed into expression systems. Thus, it may also be possible soon to prepare large amounts of the surface antigen(s) for use as a better defined, highly specific, and safe rotavirus subunit vaccine.

Other Viruses

A large number of other viral diseases are known; it should soon be possible to develop vaccines or antivirals to prevent or treat many of them. Among the most important are human papillomavirus, certain insect-borne viruses, hepatitis A virus, and adult T-cell leukemia virus.

IMPLEMENTING STRATEGIES FOR DEVELOPMENT OF NEW VACCINES AND ANTIVIRALS

Pursuing the above approaches to the design of new viral vaccines and specific antiviral drugs is a long-term program. Al

though the required principles are known, much remains to be done in the course of developing these approaches to yield practical vaccines or drugs. Their theoretical basis, however, is firm and their prospects highly promising.

Pursuit of prevention or treatment modalities should not be regarded as mutually exclusive efforts. Useful cross-fertilization occurs between the two activities. Treatment may be needed to "backstop" even a highly successful prevention effort, or it may be a more rational approach for some diseases where a target population for vaccination is presently difficult to identify.

In spite of the high likelihood of success in this area, there are several impediments to its realization. The major problems that need to be addressed are as follows.

1. The development of highly specific and potent antiviral drugs requires the collaboration of scientists in several disciplines, among them protein chemists, enzyme kineticists, biophysicists, x-ray crystallographers, organic chemists, virologists, cell biologists, pharmacologists, toxicologists, and clinicians. Not all such scientists would have to interact at the same time; but in the initial phases of the work, protein chemists, enzyme kineticists, biophysicists, and organic chemists will have to interact quite closely. A coordinated national effort is needed and should involve extensive collaboration among components of the entire scientific community. Such collaboration may come about through the cooperation of scientists in various disciplines on the same university campus, among several university campuses, within private companies, between universities and companies, and between all these groups and government scientists.
2. To capitalize on the opportunities that exist, there is a need to invigorate public-private partnerships. Private industry is concerned about the confidentiality of the results that are obtained because confidentiality is essential for the protection of patent rights, which are needed to recoup high development costs. Close, long-term collaboration between scientists in universities and their counterparts in industry would be highly desirable, but such cooperation will require very careful management and may also require the development of new mechanisms of promoting and funding it.
3. A serious impediment is the threat of possible liability for inadvertent injuries attributed to vaccines. A system providing compensation to individuals who incur untoward injury from vaccines that are correctly manufactured and administered is essential, along with some means of defining much more clearly than is currently the case the limits of manufacturers' liabilities.*
4. Another major impediment is current uncertainty about the legal limits of applicability of recombinant DNA research. A vocal minority persists in opposing any kind of innovation resulting from the application of recombinant DNA technology to human health. There is a need for more public education on the nature, benefits, risks, and practical capabilities of recombinant DNA technology. Within their mandates, agencies should attempt to provide guidance on these issues to interested parties—for example, potential manufacturers and "consumers" of products.

In the final analysis, the usefulness of antiviral drugs and vaccines should be judged on the basis of the net benefits they provide; absolute safety should not necessarily and invariably be the goal. Evaluations should take into account the number of lives saved, the misery spared, and the economic benefits accrued as well as known and potential risks.

* See *Vaccine Supply and Innovation,* a report from the Institute of Medicine, National Academy of Sciences, published by the National Academy Press, Washington, D.C., 1985.

BENEFITS OF A NATIONAL EFFORT ON VACCINES AND ANTIVIRALS

Control of poliomyelitis, measles, mumps, and rubella in the United States has produced annual savings estimated in 1980 at $2 billion. The benefits of the strategies envisaged above would be the further savings of many lives and the enormous reduction of misery and costs attributable to acute disease or to persistent, latent, and chronic infections that later cause degenerative diseases or cancer. Additional viral diseases should be eradicated following the example of smallpox.

The capabilities developed in a program on human viral diseases would also be applicable to viral diseases of livestock and poultry, where economic losses of production are enormous.

SUMMARY AND CONCLUSIONS

Recent advances in molecular virology have laid the foundation for combating many viral diseases through new vaccines or more rational approaches to the development of antiviral drugs. These new approaches utilize recent advances in the knowledge of viral surfaces and of unique processes encoded by viral nucleic acid. A central feature of the approaches to antivirals is the selection, as targets, of processes that are essential to viral multiplication but for which no host cell counterpart exists. While this is a program area in which timely addition of emphasis and support would pay great dividends, certain organizational difficulties (such as the need for large collaborative efforts) and policy issues (such as liability for vaccine injury) will need to be addressed to ensure the realization of the great health and economic benefits that these new technologies promise.

ACKNOWLEDGMENT The committee gratefully acknowledges the assistance of the following individuals in the preparation of this document: Donald S. Burke, Walter Reed Army Institute of Research; Joel M. Dalrymple, U.S. Army Institute for Infectious Diseases; Bernard Moss, National Institutes of Health; Michael Lai, University of Southern California; Stephen E. Straus, National Institutes of Health.