

linguistic theory that has an intimate and continuous tie-in between grammar "Experts" and domain-dependent "Experts". Although the domains about which they admit discourse are still modest and discrete, they are many times richer than anything done previously. The state-of-the-art is represented by the SHRDLU program for conducting a dialogue with a simulated robot about a world of blocks, boxes, and pyramids on a table; and the Lunar Rocks program for conducting a dialogue about properties of and transformations upon NASA moon-rock samples. The SHRDLU program, for example, will carry out commands, answer questions, and generally be aware of what it was doing, so as to answer "how" and "why" questions about its behavior.

The internal structure of these systems exhibits an interesting evolution over the semantic-net-memory systems, and they appear to be a long way from the heuristic search schemes mentioned earlier. They are essentially large programs written within a programming system that provides search and matching capability. There is no factorization between a data base (i.e., semantic net) and a small set of methods that process the data base. Rather, the entire system appears to be a large collection of special purpose programs for dealing with a multitude of special cases. They give the appearance of being a highly distributed system, in which the intelligent action resides throughout the entire program.

GOAL 2D. Acquiring and Understanding Sensory Data.

The goal here is to discover broadly applicable methods for extracting from sensory data (chiefly visual and aural) the information that is specifically responsive to users' needs. Two classes of needs may be noted: the need to facilitate communication between man and machine; and the need to apply computers to intrinsically perceptual tasks. The former is exemplified by the desire to talk, rather than type, to computers; the latter is illustrated by the task of automatically guiding an effector on the basis of visual data. To satisfy either (or both) of these needs, it is necessary to move from well-understood problems of sensing data to much more difficult problems of interpretation.

SUBGOAL 2D1. Visual Scene Analysis. Computer-based analysis of visual scenes has its roots in work on optical character recognition (early to mid-Fifties) and by work in automatic photoreconnaissance. These tasks are essentially two-dimensional. Little is lost by disregarding dimensions of objects in a direction orthogonal to the picture plane.

AI research on scene analysis began in the early sixties with the work of Roberts on pictures of polyhedra. This work (and its intellectual descendants) differs from the earlier two-

dimensional work in two major respects: first, it explicitly considers, and capitalizes on, the three-dimensional properties of objects and their perspective representations; second, it utilizes a variety of special processing steps and decision-making criteria, in contrast to the earlier template-match/classify paradigm.

Robert's work spawned five years of intensive research on pictures of collections of polyhedra. One theme, centered on the archetypical question "Is an edge present in a given (small) region of the picture?", led to the development of edge detecting, contour following, and region finding programs. A second theme, centered on teasing out the properties of polyhedra and their representations, led to an elegant theory of permissible representations of edges and vertices, and their relations to three-dimensional polyhedra - a theory not previously discovered by projective or descriptive geometers.

Work in the polyhedral objects domain culminated in several programs capable of describing, in more or less complete detail, pictures of complicated collections of polyhedra, even taking into account shadows cast by these objects. At the same time, more complicated types of scenes began to be seriously studied. This has led to current interest in the use of color, texture, and range data, and has stimulated interest in program organizations capable of capitalizing on these multiple perceptual modalities. For example, in one paradigm perception is viewed as a problem-solving process that uses many varieties of knowledge to select perceptual operators, to guide their application to sensory data, and to evaluate the results obtained therefrom.

SUBGOAL 2D2. Speech Understanding. Research on computer recognition of speech signal data began in the Fifties with work on the recognition of isolated words. Some observations will be made here on the relation between speech understanding research and the ongoing body of AI research.

The fundamental idea driving research on speech understanding is that "recognition" is impossible (in flowing natural speech) without understanding, and that understanding is impossible without extensive knowledge about the domain of discourse. This view arises in part from the observation that ambiguities and omissions at both the acoustic and semantic level do not arise as bizarre or pathological exceptions but instead are commonplace events. Speech understanding research thus relies heavily on progress in the basic AI research problems of knowledge acquisition, representation, and deployment. This situation is unlikely to change regardless of advances in processing acoustic signals.

GOAL 2E. Intelligent control of Effectors

This goal concerns the creation of devices and control programs for bringing about specified changes in the physical world. The effectors that have attracted the most attention have been mechanical manipulators and mobile vehicles but this has been largely a matter of experimental convenience. In principle, they could as easily have been subsystems of spaceships or manufacturing tools.

Early work in "intelligent" effectors dates back two decades, but systematic work did not begin until about 1966, at which time some progress had already been made in developing symbolic problem solving programs to control effectors. Since then there has been considerable interest in computer-controlled effectors because problems of effector control excite a set of important issues for AI research. The following is a rough characterization of the subgoals of work on effector control:

SUBGOAL E1. Monitoring Real-World Execution of Problem

Solutions: The special touchstone of effector control research is that a problem is never "solved" until the real, physical world has been altered in a fashion that satisfies the task specification (in contrast to other problem solving programs whose responsibility ends with the symbolic presentation of a good solution). Thus, an effector control program should ideally be prepared to deal with any eventuality that affects the execution of a theoretically correct solution, be it initial misinformation, accidental dynamic effects, etc. These demands strongly influence all levels of program organization and strategy. Problem solving and execution monitoring must be made to interact intimately. The most advanced work of this type is probably the STRIPS-PLANEX system (for the control of a mobile vehicle) that can detect and gracefully recover from a wide variety of execution difficulties.

SUBGOAL E2. Modelling "Everyday" worlds: To control effectors by computer requires that the computer have adequate models of everyday situations. It has become important to model occlusion, obstruction, relative location, etc., and this has been done to the extent necessary to handle various simple manipulation and locomotion problems.

SUBGOAL E3. Planning in the face of uncertainty. Problem-solving programs for the control of effectors that operate on the physical world must be able to work routinely with incomplete and

inaccurate information. This creates a need to do research on programs that can form contingency plans, can plan to acquire information, can decide when to execute actions in the physical world, even if the plan is incomplete, and so forth. Some research of this type has been done.

SUBGOAL E4. Low-Level Control. By low-level control is meant: programs that interact more-or-less directly with the effector mechanism, and that do not engage in global planning or problem solving. Research on this topic is producing a new and potentially important branch of classical automatic control. Although little has been formalized to date, enough experience has been acquired to permit the construction of interesting demonstrations. Among the most impressive of these is an arm control program that can drive the arm in partially constrained ways; for example, the arm can be made to turn a crank by dynamically constraining the necessary degrees of freedom.

SUBGOAL E5. Hardware Development. The manipulators available in 1966, whether based on prosthetic limbs or industrial put-and-take machinery, were generally too primitive to be of long-term value for AI research. This situation fostered a fairly significant hardware development effort that produced a useful arm-hand device. Similarly, sensing devices received some development efforts. Examples of this work are newly developed optical range finders, and special tactile, force, and torque sensors.

GOAL 3. Information Processing Psychology: developing detailed scientific models of human symbolic processing behavior.

Since its inception, one focus of AI research has been the study of the symbol manipulation processes capable of explaining and predicting human behavior in a wide range of cognitive tasks. As science, the endeavor is entirely classical in intent and method, employing model construction and validation. Empirical data from well-controlled laboratory experiments is obtained from psychologists or generated by the researchers in their own laboratories. Induction from this data leads to the formulation of a symbol-processing model which purports to explain the observed phenomena. This model is given a precise form as computer programs and data structures (since the computer as a general symbol-processing device is capable of carrying out any precisely specified symbol-manipulation process; this step is entirely analogous to the model-implementation step taken by the physicist when he translates his physical model into the form of a set

of differential equations). A computer is then used to generate the complex and remote consequences of the symbol-processing postulates of the model for the particular laboratory situations and stimuli being studied. These consequences and predictions are tested against empirical data; differences are noted and analyzed; the model is refined and run again; iterations continue until a satisfactory state of agreement between model's predictions and empirical data is achieved.

From one point of view, the endeavor is to be seen as Theoretical Psychology. From another point of view, it can be seen as a systematic attempt by AI research to understand intellectual activity as it occurs in nature (i.e., in humans) so that artifacts capable of performing such intellectual activity can be constructed upon the principles discovered. The interplay between these two views has been very strong.

Information Processing Psychologists have usually chosen their problems in areas that have been of "classical" concern to Psychology, though some of these areas have been reopened to serious investigation because of the successes of the information processing approaches. The following are brief sketches of some subgoals of the effort in Information Processing Psychology.

SUBGOAL 3A. Functional reasoning. Analysis and modeling has been done for human behavior in solving logic problems, complex cryptarithmic puzzles, and chess-play problems. The models, and the predictions derived from them, are so detailed that no comparison with previous work on the psychology of problem solving is meaningful. The work is a scientific revolution, and has had a great paradigmatic and methodological impact upon Psychology. The principal innovators, Newell and Simon, have had their contributions recognized by election to the National Academy of Science; Simon was awarded the Distinguished Scientific Contribution Award of the American Psychological Association, more or less the "Nobel Prize" of Psychology.

SUBGOAL 3B. Rote Memory and Short-term Memory phenomena: Storage and retrieval processes for short-term memory. Rote memorization effects. Discrimination and association learning for verbal materials. These and related phenomena of verbal learning and memory have been studied intensely by experimental psychologists in this century. A few dozen solid empirical generalizations are known. A set of closely related information processing models is capable of explaining many of these (roughly speaking, 15-20 of the "classical" phenomena).

SUBGOAL 3C. Long-term associative memory: Associative retrieval from associative memory nets of several hundred to a few thousand

symbols. Interaction of English sentence processing and memory. The symbolic representation of knowledge (i.e., facts about the world) in memory. The work is currently very active, highly promising, and is causing a mini-revolution in thinking among psychologists who study memory.

SUBGOAL 3D. Pattern induction/concept formation. Induction of models of pattern regularities in strings of symbols. Induction of the "generating rule" from the exhibition of instances of the rule.

SUBGOAL 3E. Phenomena of neurosis. The behavior studied is neurotic symbol-processing behavior, viewed as processing distortions of otherwise "normal" linguistic and problem-solving processes. A highly successful model of paranoid behavior has been developed, incorporating some English language processing.

These examples are but pieces of a bigger picture, which looks something like this:

1. It is no surprise that Psychology has been strongly affected by the information processing concepts and tools of AI research since both sciences are concerned with the study of cognition. The magnitude of the impact is the big surprise. It is probably fair to say that the dominant paradigm currently structuring Experimental Psychology in this country is the information processing paradigm. Upon no other area of science has AI research had such a strong impact.
2. The scientific study of human thought has been accelerated greatly during the last fifteen years because of the AI impact. It is not much of an overstatement to say that the AI impact has revitalized the study of thinking by Psychology, making this scientific enterprise tractable, fruitful, and respectable.

VIEW OF THE FUTURE: What lies within a five year horizon?

An extrapolation of the research directions previously described into the future faces at least two problems. First, there are the usual uncertainties that loom because of unpredictable advances and wishful thinking. Second, the imposition by ARPA of research priorities upon the course of events that would "normally" ensue will have a large effect. Thus, the question of "what should happen" is as big a question as "what will happen."

This exposition is made difficult by the fact that the structure of the field, as outlined above in terms of Goals, will show strong confluences during the future period. Any simple presentation goal-by-goal would be misleading, and was not attempted. Instead, each identifiable focus is stated and then given an extended discussion.

The main thrusts of the Artificial Intelligence community in the next five year period will be:

1. Development of applications programs that represent and use knowledge of carefully delimited portions of the real-world for high-performance problem solving, hypothesis induction, and signal data interpretation.

The next period is likely to be a period of consolidation of AI's previous gains into meaningful real-world applications. High levels of competence in the performance of difficult tasks will be the hallmark. In addition to growing attitudes toward becoming more relevant, the AI community's current major interest in knowledge structuring and use will naturally lead it to bodies of real-world knowledge that are rich in structure and challenges. An extrapolation indicates applications to domains in science (much as the DENDRAL and MATHLAB programs were developed); and in medicine (current activity includes programs that deal with Infectious Diseases and with Glaucoma); perhaps more routine aspects of architecture (e.g., space layout and design); perhaps design in electronics (e.g., layout of IC and PC electronics, actual circuit design to functional specs); management science applications (e.g., logistics management and control, crew scheduling for aircraft fleets). The most significant application will be to computer science itself, namely the automation of many programming functions (to be discussed later). Application to some of the less routine aspects of office document processing is a likely event (discussed later). With appropriate stimulus from ARPA, or other service agencies, these application priorities could be shifted toward defense problems, particularly those related to signal processing (e.g., application to seismic or sonar signal interpretation). In such applications, interpretation of what the signal means is made in terms of knowledge about the signal-generating source and the environment in which the signal occurs. All of these applications will be characterized by careful choice of domain, careful delimiting of the extent of knowledge necessary to do the job, and close coupling with human experts to gain the knowledge necessary. None of these programs will be "general problem solvers" of the old genre. Characteristic of some of these applications will be one-line interaction with human experts, not only to "tune" the knowledge used by the program, but also to intervene in decisions for which human expertise dominates that of the program, or where the relevant knowledge has not been made explicit and formalized for computer processing.

2. The development, in particular, of that area of application involving the synthesis of computer programs (the so-called "automatic programming" problem).

The particular application of AI techniques to the task of synthesizing computer programs from imprecise and non-procedural descriptions of what a user wants a computer to do for him is the AI problem area whose time has come. This area will be the subject of a separate and detailed program plan. It is an AI application of tremendous economic, and industrial importance, since computer programming is today a major bottleneck in the application of computers to technological and business problems. What is worse, virtually no advances of substantial impact upon this problem have been made in the last decade in other areas of computer science (with the possible exception of the interactive editing, debugging, and running of programs). The automatic programming problem is, furthermore, the quintessential problem that fits the WHAT-TO-HOW characterization of the nature of the science of Artificial Intelligence. It is the meeting ground of many of the tributaries of AI research: problem solving, theorem proving, heuristic knowledge and search, understanding of English (perhaps even speech), and advanced systems work. It is an ideal problem from the viewpoint of knowledge-based systems--the main line of current AI research. The essential activity in building such systems is the extraction and formalization of knowledge of the specific task domain. In the art of programming, computer scientists are their own best experts, and for years have been engaged in formalizing what is known about programming, mathematically and in other forms. Following this line of reasoning, the programming task that may be best suited is systems programming. An example of a specific systems programming task that may be accomplishable within the period is: development of an automatic programming system that will produce operating system code for a minicomputer like the PDP11/45, in response to functional specifications for instrument control and data-handling, where the specs are given in functional terms by a scientist putting together the instrument-computer package, not his (until now inevitable) programmer.

3. The extension of current ideas about the processing and understanding of English to more extensive domains of discourse and with greater flexibility, to the point of practical front-end processors for large applications programs.

In the coming period, programs for understanding English in limited "universes of discourse" will achieve practicality, and will be made available as the linguistic interaction vehicle in some of the larger AI applications programs, e.g., the

automatic programming systems mentioned above. Since these applications programs will be domain-limited anyway, it will not be an extraordinarily difficult task to construct for them front-end processors that understand English in that domain. Since currently the field has only "demonstration programs" that exhibit (limited) understanding of English, much more research will be undertaken in these directions: examining how well current techniques extrapolate to broader domains of knowledge; developing techniques for establishing context of an interaction and maintaining that context throughout the conversation; and extending methods for drawing inferences from the continually updated context. Research on semantic theory, previously mentioned in connection with representation of knowledge, will be applied to specific problems of linguistic interaction involving actors, actions, objects, and common-sense knowledge. The area of language understanding is so rich in possibilities and implications that it is not unreasonable to consider developing a separate program plan for it within the next two years.

4. Initial exploration of office-work tasks as an area of development and application; the careful choosing and shaping of specific tasks in this enormous arena of human endeavor; and some limited applications progress on these tasks.

The AI research community has been searching for problem domains of significance to science, technology, or industry that would provide an integrating theme for the various subareas of AI work. These subareas have a considerable coherence of concepts and techniques, but the centripetal force of a real-world theme is necessary to make this coherence a practical reality. Production assembly by combinations of vision, manipulation and problem solving programs is an attempt to establish such a theme. Increasingly the feeling is growing in the AI community that the development of "intelligent assistant" programs for ordinary office work is a useful and important focus. There are two reasons for this. First, much of current AI research fits the task area well (e.g., semantic-net-memory structures, question answering programs, natural language understanding, "intelligent assistant" interaction programs, etc.). Second, the explosion of use of the ARPA network for "office work" tasks quite apart from computation (uses such as message processing, message and document filing, information retrieval from large data bases, composing and editing of documents, etc.) provides an excellent medium in which to do the work. The AI community, perhaps with a push from ARPA, has the capability to do significant work on the office automation problem in the next period. A carefully thought-through program plan will probably be the first output of the field in this area (should be organized and completed within the next two years), followed by initial exploratory

ventures along the lines laid down in the plan. Again, as with all the knowledge-based systems of this decade, the specific tasks worked upon will of necessity be carefully delimited. The general "intelligent office assistant" is well beyond the horizon, but specific assistant-programs for handling some of the office-work flow of information on the ARPA network can be realized within five years.

5. Intensive developmental work on the speech-understanding problem.
6. Expansion of computer vision research to: knowledge-based program organizations; development of a repertoire of low-level perceptual operators for color, range and texture, and exploitation of these modalities; first practical applications of scene analysis to selected tasks in industrial and biomedical settings; and use of interactive scene analysis for both research and application purposes.

Scene analysis programs consist of a combination of sensing-and-measuring primitive perceptual operators (like line-finders) and higher-level knowledge-based procedures (like line-proposers). Because of general awareness of the limitations of current primitive operators (at least as they are applied to monochrome pictures), the research will place increased emphasis on the acquisition and low-level analysis of color and range data. Higher level procedures will use knowledge of: three-dimensional properties of objects other than polyhedral objects; perceptual properties of objects; many varieties of contextual constraints among objects; and properties of the primitive operators (like computational cost, reliability, and domain of applicability). Practical applications will probably focus on industrial tasks like work-piece identification and location, inspection, and manipulator control. The scene analysis research issues in these applications may turn out to be pedestrian, but concerns about cost, reliability, and reprogrammability will become prominent. Biomedical scene analysis problems will continue to stimulate research; application to medical mass-screening tasks may occur. Interactive scene analysis will be an important focus. In research settings, interactive scene analysis will be used to construct large scene-analysis systems through the incremental accumulation of knowledge; in application settings it will be used to achieve flexible scene analysis systems that can be easily "re-programmed" by users who are not computer scientists.

7. Expansion of arm-hand effector technology and associated program control, with some practical applications of simple forms of this technology in industrial settings.

There will be considerable activity in the transfer of ARPA-initiated work on effector control to industrial settings. Hardware realizations of a rich variety of mechanical effectors, with their tactile, force, and torque sensors, will appear. Visual feedback in controlling effectors will be a feature of many of the applications. Basic research on the hardware and software technology of effector control will continue, if support from ARPA or other agencies is forthcoming. More broadly-based research on effector control is likely to be stimulated by the appearance of relatively inexpensive experimental hardware. Researchers who are currently unable to develop one-of-a-kind devices because of their cost will enter the field.

8. Expanded basic research on acquisition, deployment and representation of knowledge to support knowledge-based systems development.

Though the main thrust of AI research is in the direction of knowledge-based programs, the fundamental research support for this thrust is currently thin. This is a critical "bottleneck" area of the science, since (as was pointed out earlier) it is inconceivable that the AI field will proceed from one knowledge-based program to the next painstakingly custom-crafting the knowledge/expertise necessary for high levels of performance by the programs. In the next period, the following kinds of fundamental explorations must be pursued and strongly encouraged:

- a. Additional case-study programs of hypothesis discovery and theory formation (i.e., induction programs) in domains of knowledge that are reasonably rich and complex. It is essential for the science to see some more examples that discover regularities in empirical data, and generalize over these to form sets of rules that can explain the data and predict future states. It is likely that only after more case-studies are available will AI researchers be able to organize, unify and refine their ideas concerning computer-assisted induction of knowledge.

- b. Development of interactive interrogative techniques, coupling a program to a human expert, by means of which the program systematically elicits from the expert particular facts, useful heuristics, and generalizations (or models) in the domain of the human's expertise. Again, specific case-studies are desirable. Their development need not await the arrival of English language understanding programs to facilitate the interaction and interrogation. (Stylized languages designed for the specific case-study domains will serve for now.)

c. Exploration of a variety of methods for bringing together disparate bodies of knowledge held by a program to assist in the solution of specific problems that the program is called upon to solve. The nature of this problem was discussed earlier under Goal 2A. If there are to be a number of Experts (i.e., specialized knowledge bases) interacting in the solution of a problem, how should their interaction be arranged? Is there an Executive Program "in charge" of sequencing the activity of the Experts? If so, what is the nature of the Executive Program's knowledge about each Expert, and the appropriateness of calling that Expert to assist at a specific point in the process? Should the Experts be relatively independent, each with its own situation-recognizer to trigger its activity? These particular questions are posed here not in an effort to characterize the problem completely (or even adequately), but to give the flavor of the experimental inquiry that needs to be pursued in the coming period - a period in which major AI programming efforts will be directed toward knowledge-based systems with multiple sources of knowledge.

d. Theoretical and experimental studies of representation of knowledge. This basic and difficult problem is not one that is likely to have a "solution" in a five year period. Theoretical studies will continue to search for a logical calculus in terms of which to formalize and store knowledge in a fairly "natural" way, and for logical processors that will compute efficiently within this formalism. Experimental studies will attempt to deal with the usual nonhomogeneity of representation among different bodies of knowledge directly, by programming translations of representations from one "natural" representation to another as necessary in those situations requiring communication between Experts for joint problem-solving.

9. Continuing basic research on various mathematical-logical problems such as formal models for heuristic search, theorem proving methods, and mathematical theory of computation.

Because heuristic search has been a central theme of AI problem solving research, it is likely that attempts at mathematical formulation and analysis of heuristic search methods will continue. No existing research thrusts indicate that this work should have high priority at this time. However, the situation is unstable in the sense that a few key results (e.g., new theorems or, more likely, new formulations of heuristic search) might cause a rush of activity along lines of formal analysis.

A similar situation attends theorem-proving research. There are currently no critical ideas acting as a forcing function, but nonetheless the problem appears to some scientists to be central for progress in the long run. In their view, to state

that a computer can be used as a "symbolic inference engine" is equivalent to saying that it is a "logic engine"; and what makes a "logic engine" turn over is a theorem prover over the domain of some logical calculus. The search for appropriate logical calculi and associated theorem provers will therefore continue.

The work in mathematical theory of computation has been peripheral to the AI mainstream, but recently has been gaining momentum and importance, and will enter the mainstream as basic research for automatic programming efforts. To write programs capable of synthesizing programs obviously requires a thorough understanding of the nature of programs. One kind of understanding is gained by formal description and mathematical analysis (the kind of understanding we take so much for granted in some physical sciences and engineering). To the extent that useful formal descriptions of how programs are put together and what programs do can be discovered; and to the extent that powerful theorems can be proved within the formalism; the work on mathematical theory of computation could aid significantly in the practical work of constructing automatic program synthesizers and verifiers. Thus, there are noteworthy "breakthrough" possibilities in this area.

A prediction of the most likely course of events in these tasks of formal analysis is that they will be low-key, low cost, high risk/high payoff.

10. Continuing research on modeling of human cognitive processes using information processing techniques.

At the interface between AI and the psychology of human perception and thought, the research tempo has been increasing for some time. In the coming period it is likely that new methodology, new conceptual insights, and new models will have a continuing dramatic impact on Psychology. The feedback to ongoing AI research will continue to be important, particularly in the areas of perception and memory. The principal developments are likely to be these:

- a. Methodological: analysis by program of the thinking-aloud protocols of humans solving complex problems (i.e., "data reduction" that requires some language understanding and complex inductive inference), resulting in a speed-up in this critical empirical procedure of perhaps a factor of 100. A typical complete protocol analysis of human data in a puzzle-solving task currently takes, without computer assistance, 100 hours.

- b. Short-term memory. The processes of human short-term memory will be so well modeled and understood as a result of

research in this period that the topic will cease to be of major theoretical interest to psychologists.

c. Long-term memory. A very good model of human long-term associative memory will be developed. The program which realizes this model will be given a great deal of "garden variety" knowledge of the everyday world, as the basis for empirical testing. Such a model will undoubtedly prove to be an important subsystem in larger programs that attempt language understanding in contexts involving common-sense knowledge. Only the beginnings of such a memory model exist today.

d. Visual perception. The most important impact of AI on Psychology in the coming period may be the initial formulation on an information processing theory of human visual perception of common 3-D forms, along the lines of the visual processing concepts and operators developed by AI vision research. AI vision research stands on the threshold of Psychology awaiting an intellectual push like the one given to problem solving in late Fifties. If the push is made, and is successful, it will noticeably dent the theory of visual perception in five years and totally capture it within ten years.

APPENDIX B

Justification for Storage Augmentation - July 1974

The following is the text of a proposal submitted to the AIM Executive Committee and the NIH-BRB at the end of July 1974 for augmenting file storage, memory, and swapping storage capacities for the SUMEX-AIM resource. The committee approved the proposal and, as discussed in the text of the report, we have implemented the file space and memory additions to good effect. The swapping storage augmentation has been pending until we felt a clear demonstration of need existed.

Based on system performance measurements over the past several months, we have come to the belief that whereas we are at the capacity of the swapping storage now under peak load, there may be a software remedy which will delay the need in this area. SUMEX CPU capacity has become a rather more critical resource at this time with the growing community of users. We are currently formulating an additional plan to augment the system in terms of processor power. This will be submitted for review during the next grant year (03).

RECOMMENDED SYSTEM STORAGE AUGMENTATIONS
WITHIN FIRST YEAR BUDGET
JULY 29, 1974

The initial SUMEX computer configuration plan, approved by the AIM Executive Committee in November 1973, was a compromise between the technical demands of establishing an effective community AI computing facility and the budget constraints imposed by Council. Within the projected budget at that time, we attempted to balance the configuration in terms of available file space, core size, and swapping storage.

As discussed at earlier Executive Committee meetings, ARPA has found it necessary over the past 6 months to reconsider its policies as they relate to ARPANET expansion and use by non-DOD agencies. The result of these deliberations has been a decision in early July by ARPA that SUMEX can become a Very Distant Host (VDH) on the ARPANET rather than a new TIP node as initially planned. We have revised the earlier network plan to implement a VDH interface and to augment the interim line scanner capacity to handle local terminals (previously to be handled by the TIP). We are also in the process of interfacing to the TYMNET in order to provide low bandwidth terminal support on a broader geographical and administrative basis than is afforded by the ARPANET at the present time.

Some reductions in first year costs have resulted from the reconfiguration and delays in implementing the ARPANET connection. These include delayed project staffing, delayed operational status, and reduced communications fees as well as the inherently lower cost of the VDH connection. The overall reductions amount to approximately \$148,000 and afford the opportunity to reconsider other aspects of the machine configuration to give a larger capacity to better meet the needs of the AIM community.

Whereas the SUMEX facility is just coming to a fully operational state, we can project a number of areas where augmentation would be of benefit to system performance. These projections are based on observations of current SUMEX utilization as well as experiments on a KA-TENEX system at the Institute for Mathematical Studies in the Social Sciences (IMSSS). The IMSSS machine allows a more parametric measurement of performance sensitivity to hardware changes because it has a larger configuration from which the effects of reducing various component capacities can be observed. The following summarizes these recommendations.

FILE SPACE

Even in these early stages of SUMEX operation, it has become clear that the file system capacity will be a limiting factor to AIM community expansion. This derives from the interactive nature of the TENEX system making on-line files essential, the large files involved in AI program images, and the large data files currently in use and expected increasingly as data base-oriented AIM projects are identified. The capacity of the current file system is not yet fully utilized and we have issued only verbal requests to economize on file space. However, the trend toward early consumption of the file capacity is clear as summarized by recent file utilization statistics.

Out of a total of 81,200 available pages (4 RP-03 disk drives), the following are averages of the space in use including all system and user directories:

Mid-June	47,500 pages
Late June/early July	53,000 pages
Mid-July	52,000 pages
Late July	52,000 pages

We have developed a policy statement on file space allocation and control which is attached. In this policy, current data on disk requirements for various aspects of the system and user projects are integrated to allocate the overall available space (81,200 pages):

I. TENEX/AIM SYSTEM (common to both SUMEX-SUMC and -AIM)

Operating Monitor	5,000 pages	
Supporting Direct. (lang., lib., etc.)	10,000 pages	
AIM management and SUMEX staff	10,000 pages	
File system reserve for temporary overflows	6,200 pages	
	TOTAL	31,200 pages

II. SUMEX-SUMC Users

TOTAL	25,000 pages
-------	--------------

III. SUMEX-AIM Users

TOTAL	25,000 pages
-------	--------------

81,200 pages

Among the initial SUMEX-SUMC projects (DENDRAL, Protein Structure Modelling, MYCIN, and various pilot efforts) approximately 17,500 pages are in use. On the SUMEX-AIM side only 8,000 pages are allocated because delays in network connections have precluded Dr. Amarel's and Dr. Colby's groups from actively using the system.

Based on these data, we recommend adding 4 more drives (81,200 pages - this is also the limit of the number of drives which can be put on the existing controller) to augment the SUMEX-AIM component of the file system. This would provide room for an additional 8-16 projects at 5,000-10,000 pages per project. At \$13,000 (plus tax) per drive, the total cost for this augmentation would be \$55,120.

MEMORY AND SWAPPING STORAGE

The operational status of the SUMEX KI-TENEX system has been approaching "routine" since May for the local community primarily. Over this period we have begun to collect statistics on the performance of the system but note that swapping is implemented on a provisional, inherently inefficient basis on the moving head file system disks. A sample of these data is shown in Figure 1. During the prime time shown, the system load was 10-14 jobs including 2 or 3 LISP users and miscellaneous EXEC, editor, and private program jobs. Plots are shown in Figure 1 of the percent time allocated to running user programs and the percent time consumed in system overhead (waiting for pages to be swapped in and out to make a program runnable, managing core allocations, and handling page fault traps). It is significant to note that the overhead consumes on the average about 35% of the machine under this load and in excess of 60% at times. This is predominantly a result of I/O waits on the relatively slow disks used for swapping. During this period, the maximum demand for swapping storage was 1750 pages.

A dramatic improvement in efficiency is expected when our permanent fixed head swapping device is installed in August, but these data raise obvious questions about the system capacity which will be allocatable to additional user projects. In conjunction with Mr. Rainer Schulz of the Stanford IMSSS facility, we have collected a preliminary set of data illustrating the relationship between system overhead and hardware configuration. The IMSSS KA-TENEX facility was used because they have a total configuration of 256 K words of memory and a large swapping drum in operation so that by limiting each of these parameters, we could evaluate the overhead under a "standard" load. The results of this experiment are shown in Table 1.

At present, the SUMEX machine is operating in a configuration similar to box 5 in Table 1 and with the installation of the swapping device will operate somewhere between boxes 1 and 3. (Note that the amount of virtual address space overflowing the "drum" determines the relationship of box 3 between boxes 1 and 5). The interaction between overhead, configuration, and job mix is complex. Witness for example, some data not shown in the Table. By adding 2 100 page jobs to the 4 200 page jobs in boxes 3 and 4, the overhead in box 3 is lowered while that in box 4 is raised. Nevertheless, several general relative trends can be noted. Increasing the speed of swapping storage reduces system overhead by reducing the I/O wait time for moving pages in and out of memory. Increasing memory size also reduces the overhead by allowing more working sets to be resident simultaneously thereby giving more candidate jobs to be run while waiting for pages to be swapped for other jobs.

It must be noted that the jobs run in this test simulate

the effects of simultaneous very large jobs. In general there will be a spectrum of job sizes which will tend to reduce the overhead in all configurations (more working sets resident). On the other hand, the overhead estimates for swapping off of moving head disks are low because no data files were in use during the test thereby necessitating fewer time-consuming head seeks than would be encountered normally. Also the test programs addressed their arrays sequentially so that large blocks of pages would tend to be sequentially resident on disk. Thus in swapping programs in and out, less seeking would be required than normal.

From these estimates of relative system overhead as a function of configuration, it is clear that substantial gains can be made by adding memory to the system and by guaranteeing enough capacity so that paging occurs off a fast, fixed head device. This relative overhead can be reduced from something in excess of 30% (box 3) to something in excess of 11% (box 4) by adding memory and from greater than 11% to about 8% by adding more swapping storage. The improvement in efficiency by adding swapping storage would in fact be more than is apparent from the above data, taking into account the additional inefficiencies involved in more randomized disk seeks. Note that on the day data were taken for Figure 1, the maximum swapping space in use was 1750 pages. The fixed head swapping device we are getting will have a capacity of 2600 pages. Thus, in normal operating circumstances the probability that swapping storage will overflow to the slower moving head disk is real.

Even for a 100% efficient system, the number of users which can effectively be accommodated is limited by the response time for each user given roughly by a subdivision of the CPU capacity between the total number of users. It is very hard to pin down this number at present because it will depend on the nature of the jobs in execution. In the grossest terms, we might expect one limiting complement of users to be on the order of 5-10 LISP jobs (300-400 pages each) and 20 smaller jobs (50-100 pages each) for a total of something over 4000 pages of address space in use. This would clearly overflow the 2600 page swapping device.

For the above reasons, even though the firm limits of the current machine configuration have not been reached by existing user community demands, augmentations of the system memory and swapping storage would be beneficial to the AIM mission in allowing a larger community of projects to participate. Within the first year budget allocation, 64 K words of fast memory can be added (\$50,000 plus tax) and the swapping storage doubled (\$37,600 plus tax). Based on the relative data in Table 1, these additions, while costing about 10% of the overall facility, may free up approximately 20% of the machine capacity from overhead. This extra capacity is significant in terms of added AIM user

support. We therefore recommend these augmentations in addition to the file system expansion discussed previously.

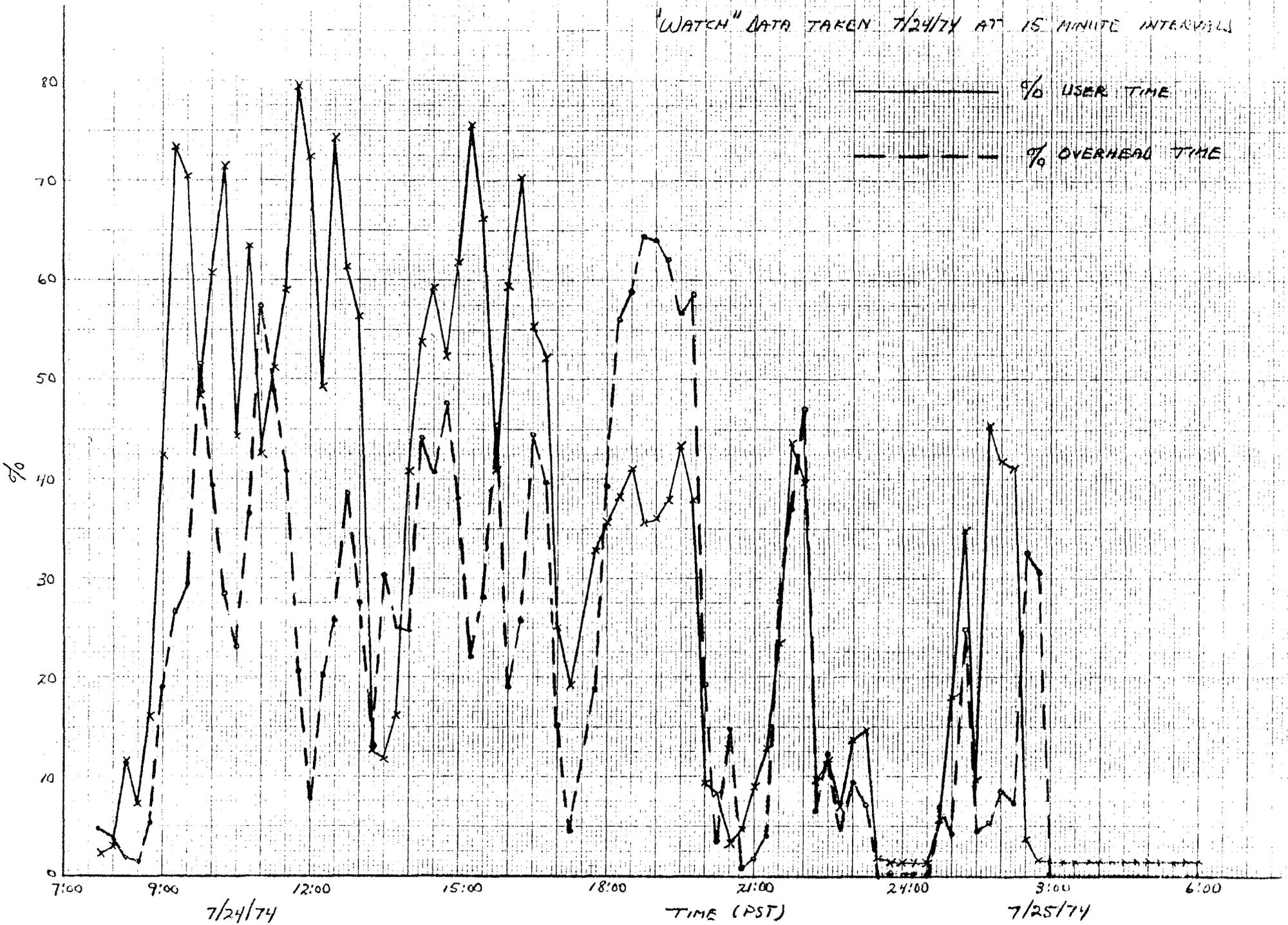
The total augmentations can be accommodated within the expected first year budget underrun:

File storage	\$55,120
Memory	\$53,000
Swapping storage	\$39,850

TOTAL	\$147,970

FIGURE 1

"WATCH" DATA TAKEN 7/24/74 AT 15 MINUTE INTERVALS



141

Table 1

System Overhead as a Function of Configuration

	Memory	
	196 K	256 K
All "drum" (fixed head)	1 9 x 100 pages: 6% 4 x 200 pages: 16%	2 9 x 100 pages: 4% 4 x 200 pages: 8% 4 x 300 pages: 24%
Part "drum" / Part "disk" **	3 9 x 100 pages: (11%)* 4 x 200 pages: (33%)*	4 9 x 100 pages: (5%)* 4 x 200 pages: 11%
All "disk" (moving head)	5 9 x 100 pages: 17% 4 x 200 pages: 51%	6 9 x 100 pages: 6% 4 x 200 pages: 15%

* Estimated by interpolation because actual measurements were not available

** The drum space was limited to 450 pages with any overflow moving to disk

APPENDIX C

Assessment of System Responsiveness Under Load

The reports from the individual projects in the SUMEX-AIM community (see Section IV, for example page 80 and page 93) suggest that the system loading is subjectively approaching saturation and express concern over the ability to work during prime time and to be able to have physicians use the interactive programs with enough responsiveness so that their frustration does not go so high as to discourage them from further use. In addition to these comments volunteered in the reports, we have asked other users as well to gauge their subjective impressions of responsiveness and those of their medical collaborators against load average measurements. Most express concern about being too precise in their judgements but generally agree that very noticeable response degradations set in when the load average gets above about 4 or 5 and that responsiveness deteriorates increasingly (non-linearly) above that. In several instances users expressed concern about the long time needed to load (not just execute) large programs when the system is heavily used. Still others get frustrated when they seemingly get no attention from the CPU at all during some intervals of time. A table relating subjective feeling to load average as submitted by one user is reproduced below as being fairly typical of the reactions received.

" LOAD AVERAGE -----	PERSONAL FEELING -----
< 1.0	Heaven. Echo great, no delay. interactive programs are. When response is not immediate, you know something important is happening.
1.0 - 2.0	Not bad at all. Slowdown is perceptible, but easily tolerable.
2.0 - 3.0	Livable. Echos are delayed by now, and impatience begins on waiting for typeout to catch up. Simple problems in a LISP job are doable, but larger ones are getting long.
3.0 - 5.0	Only editing proceeds with ease. I do practical problems only if they are extremely important (this gives me an increasingly small window these days, as the load average is frequently in this area).
> 5.0	Interactive programs aren't very, and aren't at all much above 5.0. The most trivial of demos is painful, as the response (echo) can be seconds, and the program response on even a simple task is long. Even FORTRAN programs bog down badly here. "

It is difficult to precisely quantitate the subjective aspects of response time, relating user frustrations to objective loading measures, because of wide variations between user personalities and the interactive quality of various programs. Typically, the more intimately interactive a program is the more easily user frustrations are built up with response delays. The fairly commonly expressed break point in the responsiveness versus load average curve at a load average of 4 is understandable as this is about the number of runnable working sets which can be kept in memory at a given time (user memory is about 380 pages which will hold 4 working sets averaging 95 pages each). With fewer than 4 or 5 active jobs, each job gets an aliquot of CPU time periodically on a fairly continuous basis. The main effect causing slowness is that a given user gets only about one quarter or one fifth of the PDP-10. Above this level, some jobs have to be swapped out and get no CPU time until when they are again brought in. From the user's point of view, the system sits idle on his job for an interval of time and then he gets a interval of attention. These intervals of inactivity produce particular frustration for some people as indicated in some of the comments. The problem is especially acute for large jobs (LISP programs) because they are more likely to have to be swapped out. Smaller jobs (like text editors, some language processors, and utility programs) tend to fit into the "cracks" when memory is allocated and hence see better service. This situation is reflected in the user comments as well.