FOUNDATIONS OF COGNITIVE SCIENCE: OVERVIEW

Technical Report AIP - 46

Herbert A. Simon & Craig Kaplan

Department of Psychology
Carnegie Mellon University
Pittsburgh, Pa. 15213

June 1988

This research was supported by the Computer Sciences Division, Office of Naval Research and DARPA under Contract Number N00014-86-K-0678; and by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionic Laboratory under Contract F-33615-78C-1151. We thank our colleague, Patricia A Carpenter for helpful comments and suggestions. Reproduction in whole or in part is permitted for purposes of the United States Government. Approved for public release; distribution unlimited.

Foundations of Cognitive Science: Overview

Herbert A. Simon and Craig Kaplan
Carnegie-Mellon University

[1] This chapter, billed as an "overview," will treat rather concisely a range of topics that contribute to defining the scope of cognitive science and the numerous dimensions along which this field can be explored. It will focus more on identifying issues than on providing definitive answers, a task more appropriately left for the chapters on specific topics.

Since many different methods are employed in cognitive science research, since several different conceptual approaches have been taken to the field, and since a number of different architectures for intelligent systems are being explored, we will take some pains to recognize this diversity (which should not be mistaken for confusion). At the same time, we will not hesitate to express opinions about which ideas seem most correct or promising, and which lines of inquiry seem most worthy of pursuit. The authors of other chapters will have ample opportunity to redress our imbalances.

1. The Goals of Cognitive Science

Cognitive science is the study of intelligence and intelligent systems, with particular reference to intelligent behavior as computation. Although no really satisfactory intensional definition of intelligence has been proposed, we are ordinarily willing to judge when intelligence is being exhibited by our fellow human beings. We

[1] This research was supported by the Personnel and Training Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-86-K-0768; and by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-78-C-1151. Reproduction in whole or in part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited. We thank our colleague, Patricia A. Carpenter for helpful comments and suggestions.

say that people are behaving intelligently when they choose courses of action that are relevant to achieving their goals, when they reply coherently and appropriately to questions that are put to them, when they solve problems of lesser or greater difficulty, or when they create or design something useful or beautiful or novel. The reason for applying a single term, "intelligence," to this diverse set of activities is that we expect that a common set of underlying processes is implicated in performing all of them.

When we wish to compare different people on a scale of intelligence, we construct batteries of tests presenting a variety of tasks that require them to solve problems or to use language appropriately. There are innumerable kinds of tests that exercise the intellectual capabilities of the test takers. Some of them call upon knowledge of very specific subject matter, but those we specifically label "intelligence tests" are usually designed to be as independent of subject-matter knowledge as possible -- or at least only to draw upon subject matter that is presumed to be familiar to most members of the culture (general vocabulary, arithmetic, and the like). In this chapter we will discuss both knowledge-based and (almost) knowledge-free intelligence.

Today, it is quite common to attribute intelligence both to human and to non-human systems, and in particular, to programmed computers. Not everyone accepts this usage, but we will call programs intelligent if they exhibit behaviors that would be regarded as intelligent if exhibited by human beings. Intelligence is to be judged by ability to perform intellectual tasks, independently of the nature of the physical system that exhibits this ability.

Cognitive science, defined as the study of intelligence and its computational processes, can be approached in several ways. We can undertake to construct an abstract theory of intelligent processes, divorced from specific physical or biological implementations. We can study human (or animal) intelligence, seeking to abstract a theory of intelligent processes from the behavior of intelligent organisms. Or we can study computer intelligence, trying to learn the computational principles that underlie the organization and behavior of intelligent programs.

In fact, cognitive science follows all of these paths. Several venerable examples of the study of intelligence in the abstract predate the computer. Formal logic is one such example; the theory of maximization of expected utility is another. For a century or more, experimental psychology has been studying organismic intelligence, especially as it is exhibited by people, rats, and pigeons in the laboratory. Since at least 1950 (we might take Turing's essay on "Computing Machinery and Intelligence" as a convenient starting point), that branch of computer science called "artificial intelligence" has been studying the intelligence exhibited by machines. (Watt's governor, or Pascal's calculating machines, might be regarded as even earlier examples of machine intelligence.)

For the purposes of this chapter, then, we will define cognitive science as the study of intelligence and its computational processes in men (and animals), in computers, and in the abstract. It will be instructive to see how the communality among these three topics came to be recognized, and how that recognition led to the birth of the discipline of cognitive science.

2. The Principal Contributing Disciplines

From a sociological standpoint, disciplines are defined less by their intellectual structure and content than by the scientists who identify with them. More accurately and in the long run, the intellectual content of a discipline gradually defines its boundaries and membership, while its membership gradually redefines its content. Therefore, it is important, if we are to understand cognitive science, to know what disciplines have contributed to its formation. Among these, we must certainly count experimental and cognitive psychology, artificial intelligence (within computer science), linguistics, philosophy (especially logic and epistemology), neuroscience, and some others (anthropology, economics, and social psychology will come in for comment presently).

Psychology

From its beginnings, the discipline of psychology has been concerned with intelligence. The Binet-Simon intelligence test (the "IQ" test par excellence) dates from the turn of the century. However, the dominance of behaviorism during the first half of the century prevented experimental psychologists from being much interested in what was going on inside the organism, hence limited speculation and research about process. Brain research contributed to our knowledge of the location of functions within the brain, but had relatively little to say about cognitive processes. Even the precise physiological basis of memory was not (and has not yet been) unambiguously determined.

During the high tide of behaviorism, experimental psychology focused upon relatively simple cognitive performances, with emphasis upon sensory and motor processes: rote verbal learning, tracking tasks requiring eye-hand coordination, memory tasks involving relatively short-term retention, and the attainment of simple concepts are typical examples. The intelligence of rats and pigeons received as much attention as the intelligence of people. It was left primarily to the Gestalt psychologists to develop theories of human cognitive processes, especially for complex cognitive performances like concept formation and problem solving.

Psychology was no more monolithic than chemistry or physics: various specializations were (and are) visible, each of which has brought its particular contribution to cognitive science. Psychometrics brought its measures of intelligence and the components of intelligence. Neurophysiology brought knowledge of the biological structures that support thought. Experimental psychology brought a host of information about the speed and limitations of simple sensory, perceptual, motor and memory processes. Gestalt psychology brought hypotheses about the processes that occurred in complex thinking. Although there was some communication among these specialties, each tended to go its own way, guided by its own paradigm. A new paradigm was required to make a convincing case for their mutual relevance.

The shift came with the so-called information processing revolution of the 50's and 60's, which viewed thinking as a symbol manipulating process, and used computer simulation as a way to build theories of thinking. A relatively new specialization, psycholinguistics, found the information-processing view congenial, and opened up a line of communication between psychology and linguistics.

Main-line experimental psychology also began to adopt the information processing point of view, without necessarily embracing computer simulation, and a rough division of labor began to develop between those scientists (often computer simulators) who studied the so-called "higher mental functions," like concept formation, problem solving, and and use of language, and those (less often computer simulators) who studied simpler, and more traditional, memory and perceptual tasks. The chapters by Simon (1979) and Posner and McLeod (1982) in the *Annual Reviews of Psychology* provide a good overview of these two parts of the cognitive psychology scene in the late '70s and early '80s.

Artificial Intelligence

The very term "artificial intelligence," coined about 1956, incorporated the belief that the concept of intelligence now had to be extended beyond human and animal performance to include artificial systems -- computers. Since it is by no means evident just what the intelligence of people and the intelligence of computers have in common, beyond their ability to perform certain tasks, the close relation that has been maintained between research in AI and research in cognitive psychology was not preordained, or even predictable.

The earliest artificial intelligence programs (e.g., the Logic Theorist (Newell & Simon, 1956)) are perhaps best viewed as models of abstract intelligence; but none the less, their design was informed by psychological research on memory and problem solving -- note, for example, the use of associative structures in list-processing programming languages, and subsequently the frequent use of means-ends analysis for inference.

In turn, AI research has made numerous contributions to cognitive psychology. For example, AI provided list processing languages that permit the modeling of elaborate associative structures -- schemas, scripts, frames -- to simulate important properties of human semantic memory. AI research also adapted programming languages built up from so-called "productions" as a sophisticated interpretation and augmentation of classical stimulus-response relations and stimulus recognition processes. Research on robotics has often turned to sensory and perceptual psychology for ideas about processing schemes, while the psychology of vision and speech recognition has borrowed many ideas from AI.

In fact, then, there has been a close continuing relation between AI and cognitive simulation during the whole thirty-odd years of the history of both subjects, and their mutual relevance and synergy was a major motivation for creating a common meeting ground in cognitive science.

Linguistics

The study of language has a long and complex history. In academia, linguistics has bases in departments of classics, of biblical studies, of modern languages, and of anthropology. Until recently, it had only tenuous relations with psychology, and today it is represented in cognitive science mainly under the labels of "computational linguistics" and "psycholinguistics."

Computational linguistics, as its name implies, is concerned with the use of computers to process language, for example in parsing and translation algorithms. Within cognitive science, it is most closely linked to artificial intelligence, although that linkage is of relatively recent origin. Early research on machine translation had only weak ties to either mainstream AI or mainstream linguistics.

Psycholinguistics, the psychological study of language, followed a partially autonomous path to contemporary cognitive science. A sizeable gulf in communications still exists between cognitive scientists who entered the field from AI or from the study of problem solving and concept forming behavior, on the one side, and those who entered from a concern with language, on the other. Newell and Simon's *Human Problem Solving* is a typical example of the former viewpoint, Johnson-Laird and Miller's *Language and Perception* and example of the latter.

As we shall see, it takes rather careful inspection to verify that both the group that focuses on problem solving and the one that focuses on language are interested in the same process: human thinking. When linguistics is approached from a computational standpoint, the relation between the two becomes clearer. When the uniqueness of language processing as a human faculty is emphasized, as it has been by Chomsky (1965, pp. 47-59) and others, the gulf becomes wider.

Philosophy

We have already pointed to logic as a long-established domain for the study of abstract intelligence. The mathematization of logic at the turn of the 20th century also showed how inference processes could be viewed as symbol manipulation, and thereby provided some of the foundation ideas for the information processing revolution. Recently, the requirements of large data bases for AI systems have stimulated new developments in modal and nonmonotonic logics for reasoning about the information stored in these bases.

In economics and statistics, the 20th century also saw the rapid development of a formal theory of "right reason" in the form of the theories of utility maximization and of Bayesian inference. We can regard these models of rationality from logic, economics, and statistics as forming an important part of a normative theory of intelligence, hence of cognitive science, although the researchers responsible for modern decision theory have had little involvement (or recognition) in the cognitive science community. The relevance of this kind of decision theory to psychology is debatable.

The relation of epistemology to cognitive science is rather different. The advent of machine intelligence forced radical reconsideration of the mind-body problem. Those who rejected the idea that machines could think (e.g., Searle, 1980) had to forge new theories to demonstrate why the appearance of intelligence was not intelligence. Those who accepted the idea of machine thinking (e.g., Turing, 1950) had to establish operational criteria, like the Turing Test, to verify the presence of intelligence.

Computer simulation methods also introduced a new precision into epistemological discussion. Arguments that were often unavoidably vague when expressed in natural language lost much of their ambiguity when rephrased in terms of the behavior of actual or potential machines or computer programs.

Neuroscience

Neurophysiology, and neuroscience generally, also occupies a very complex place in cognitive science. There remains today a large gap between reduction-in-principle and reduction-in-practice. Probably an overwhelming majority of psychologists believe that the processes of thinking are, in principle, explainable in terms of the electro-chemical processes of the brain. Many, however, believe that theories of intermediate level -- theories that take information processes and not neurological processes as their primitives -- are absolutely essential to the understanding of human thinking. Just as biochemistry is not "simply" physics, but must be pursued independently; so, in this view, thinking is not "simply" neurophysiology, but calls for levels of theory that can only be linked to neurophysiology through a sequence of connecting theories, most of which have not yet been fashioned.

Apart from this question of the relation between information processing and direct neurological explanation of thinking, neurophysiology plays a second role in cognitive science -- particularly in providing hypotheses about fruitful architectures for machine intelligence (and presumably also human intelligence). Neurophysiology has, in particular, provided a strong stimulus for the connexionist architectures that we will discuss later.

Other Contributing Fields

The fields discussed so far all have members who are affiliated to some degree with cognitive science (e.g., belong to the Cognitive Science Society). The same is true of anthropology, possibly because of the strong representation of linguistics within anthropology. Cognitive social psychology has been more remote, with its own research agenda focused on interpersonal perception, attribution, and self perception (but see Abelson, 1963, and D'Andrade's chapter in this volume).

We have seen that economics and statistical decision theory, although their central preoccupation is with rational (intelligent?) choice, have not participated notably in the developments within cognitive science (nor are they represented in this volume). Yet it would appear that their theories are quite relevant to the characterization of abstract intelligence, and vice versa. One of us has argued elsewhere that the reluctance of economists and statisticians to deal systematically with the computational limits on human and machine intelligence largely accounts for their indifference to cognitive science (Simon, 1976).

3. The Architecture of Intelligent Systems

The fundamental design specifications of an information processing system are called its *architecture.* The components of the architecture represent the underlying physical structures, but only in an abstract way. For example, an architecture for modeling the human brain might contain neurons as components, but the neurons might be characterized quite grossly as binary on-off elements with certain switching speeds. Another architecture might characterize the brain in an even more aggregate fashion, with units like long-term memory, short-term memory, sensory organs, and so on. The amount of detail incorporated in an architecture depends on what questions it seeks to answer, as well as how the system under study is actually structured.

Levels of Abstraction

The notion that architectures may be specified at different levels is best seen in digital computers. We speak of a computer as having von Neumann architecture when it has an addressable memory, capable of storing both program and data, input and output devices, and symbol processing capabilities that operate serially, including operators for comparing symbol structures and branching. The specification does not say anything at all about the physical devices realizing this scheme; and as we know from the past four decades of computing, they may be of the most varied and disparate kinds.

At the next level of abstraction, the architecture of a system may be described by defining a specific language. For example, the LISP programming language defines an architecture for a list-processing system. Memory is organized in associative structures, lists and property lists, and the basic operators of the language allow one to create and destroy lists, find an item on a list or property list, insert an item in a list, and so on. The elements on lists are symbols, themselves capable of indexing (denoting) other symbols or list structures of symbols.

In the contemporary practice of cognitive science, models of the human nervous system are usually defined at one of two different levels, called *connexionist* and *symbolic*, respectively. The elements of connexionist systems may be conceptualized as highly simplified and schematized neurons, interconnected in a network. (Not all practitioners of connexionism regard their elements as schematic neurons. Some would describe them more neutrally as representing "features.") In connexionist systems, the operators modify the network, and in particular, modify the strengths of the connections between elements.

The elements of symbolic systems may be conceptualized as symbols held in one or more memories. The symbols are usually assumed to be stored in associative structures, like those implemented by list processing languages. In some architectures (e.g., in Quillian's (1966) proposal for semantic memory, or in the ACT* system, (Anderson, 1983)), memories may be organized in networks, as in connexionist architectures; but the elements in these networks are interpreted as symbols rather than neurons or perceptual features. The symbolic network architecture describes the system at a higher, more abstract, level than does the connexionist architecture.

If we adopt a "levels" hypothesis about the architecture of the human cognitive system, then we might postulate a basic neural level, a highly parallel connexionist level above the neural level, and a symbolic level, constrained in its processing by the bottleneck of attention, above the connexionist level. Some work, but not very much, has been done to explain how symbol structures could emerge at the top level from the equilibrating activities of elements at the connexionist level.

In enumerating levels, we must keep in mind also the task environment to which a system's intelligence is to be applied. Thus, in symbolic architectures that make use of heuristic search, we have the level of the problem space in which the search is carried out, but above it, the level of the task domain that poses the problem to be solved. The problem representation that the system constructs and employs (its problem space) may be very different from the actual external problem situation (called by Newell (1980) the "knowledge level") from which the problem space is derived.

In evaluating the plausibility of different architectures as reasonable models of the human nervous system, it is well to keep in mind some of the parameters a system must fit if it is to claim that it describes human cognition. It takes about one millisecond for a signal to cross the synapse between two neurons, and longer, of course, for a sequence of such transmissions. A simple act of recognition takes roughly a second -- about one thousand milliseconds. Hence all of the activities between simple neural events and overt behaviors emanating from a few elementary information processes must be squeezed into a time range of only three orders of magnitude. Stated otherwise, an act of recognition cannot require more than a thousand successive synaptic crossings.

When the concern is with modeling the human brain, these parameters put important constraints on the architecture of a serial system. But they also put severe constraints upon the parallel systems that have been proposed, because the equlibration processes used by the connexionist systems require numerous rounds of successive approximation, or "settling down." It is not evident, one way or the other, whether their time requirements are greater or less than those of a serial recognition system like EPAM (Elementary Perceiver and Memorizer), a computer simulation that learns to recognize stimuli, using a discrimination net that performs a sequence of tests to sort them (Feigenbaum and Simon, 1986).

The "Standard" Model

For at least the two decades from about 1960 to 1980, there existed a broad (and approximate) consensus among psychologists taking the information processing point of view about the general architecture of the human cognitive system. There were many contributions to this "standard" model of cognition, beginning with George Miller's (1956) characterization of short-term memory with its seven chunks, and Broadbent's (1958) description of the memory system and its connections with sensory inputs. There is no "official" version of the standard model, but one description of it that lies pretty well within the area of agreement can be found in Chapter 14 of Newell & Simon (1972).

In the standard model, there are two principal memories, a short-term memory and a long-term memory. The short-term memory is characterized by rapid access (a few hundred milliseconds) and very limited capacity (alternatively, very short retention time). The capacity has been measured as about seven *chunks*, where a chunk is anything that has become familiarized by previous learning. Thus, for the readers of this book, English words are chunks. In cognitive network architectures, the short-term memory is simply the portion of long-term memory that is currently activated; the two memories are not separate structures.

The long-term memory is characterized by an associative organization and a virtually unlimited capacity. The time to store a new chunk is of the order of magnitude of eight or ten seconds, most of that time being used to store the information that "indexes" the chunk, permitting it to be recognized in a stimulus and accessed. The time to retrieve a chunk is of the order of two seconds for a single item, and a few hundred milliseconds for succeeding items.

In some versions of the standard model, long-term memory is assumed to contain several specialized components. For example, one component may contain declarative information, another procedural information (operators). The discrimination net or "index" to the memory may be viewed as a third component. In some versions of the model, also, the elements in long-term memory can be primed by stimulation, with the result of facilitating associative access to these elements and to elements linked with them associatively (Anderson, 1983).

In the standard model, the sensory organs are not connected directly with the main memories, but are buffered by small modality-specific memories. In the case of the eyes, the buffer is the so-called iconic memory, revealed by Sperling's classic experiment. In the case of the ears, the buffer is the "echo box," which holds auditory images for a matter of seconds. The buffers play only a limited role in most psychological phenomena, and are often omitted from descriptions of the system.

In reading most of the chapters of this book that are concerned with human intelligence, except for those dealing with vision and motor control, and those that proceed from a connexionist viewpoint, the reader will not be badly misled by taking the standard model as a first approximation to the architecture. Of course, authors of individual chapters will have their own views as to how that model must be modified and adapted to fit the facts.

Schemas and Productions

As we have just seen, information can be held in memories in either declarative or procedural form, or both. Declarative memories may be organized in schemas, each containing information (perhaps stored in list structures and property lists) about some class of structures or about particular structures. Schemas can form parts of other schemas; and conversely, schemas can contain schemas. They may be proposition-like, picture-like, or both. They may also be accessed through some kind of discrimination net (EPAM net, "index"), so that lengthy searches through memory are not required to find relevant information.

Information in procedural form can operate actively on memory. In many architectures, the operators are represented as *productions.* A production, often written C ---> A, consists of two parts: the *conditions* and the *actions.* The conditions are tests that are applied to stored structures (say, to symbols held in a particular memory). Whenever all the tests in a production are satisfied, the actions of that production are executed. These actions may alter or destroy symbol structures, or create new ones.

A *control structure*, about which we will have more to say in a moment, resolves conflicts between productions when the conditions of more than one are satisfied simultaneously. Of course, in a parallel system, all the productions whose conditions are satisfied can execute simultaneously. Productions, like schemas, can be indexed by means of a network, similar to an EPAM net, for discriminating among different sets of conditions.

Psychological theories of the associationist and behaviorist schools postulate connections (S-R associations) between stimuli and responses. The Würzburg school of Ach and his followers made use of more elaborate "directed associations" between the stimulus and task (Aufgabe), on the one hand, and the response, on the other (S + A --> R). Productions bear a certain resemblance to both of these constructs, relating a (sometimes complex) set of conditions, possibly including a goal (the Aufgabe?) to one or more actions that occur whenever these conditions are satisfied. Productions may therefore be regarded as descendants of these classical ideas, descendants that explicate and operationalize them, and implement them computationally.

Control Structures

In addition to specifying the organization of memory and the elementary processes that operate on memory, an architecture generally specifies a *control structure*, which determines the conditions under which particular operators will fire. In traditional programming languages, control proceeds from one instruction to the next, except where the execution of a conditional branching operation changes the order.

When a language provides for closed subroutines, as many modern programming languages do, control resides, at any given moment with a particular routine until it relinquishes it, either by calling on one of its subroutines or by completing execution and returning control to the routine that called it. This arrangement leads to problems of control that seem not to match very well the phenomena observed in human cognition. There is no way in which higher level routines, with a broader view of the situation than their subroutines, can force the latter to relinquish control before they finish executing. On the other hand, it is sometimes difficult to transmit from lower-level subroutines to top-level routines information that would enable the latter to exercise control in an appropriate manner.

Modern architectures for cognitive simulation seek to avoid these difficulties in various ways. The two classes of control structure that provide the main alternatives to subroutine hierarchies either use production systems, or use networks with links of variable strength.

In early production systems, productions were sometimes ordered, and conflicts (when conditions of two or more productions were simultaneously satisfied) were resolved by executing the production that came first in the order. By itself, this scheme was not often satisfactory. It was usually suplemented by a rule of refractoriness, so that a production, once executed, would not execute repeatedly with the same arguments. Another rule sometimes used gave preference to matching the memory elements that were created most recently. Another gave preference to productions with strong conditions over those with weaker conditions.

To secure a still greater measure of task-relevant control, *goal symbols* were sometimes employed in production systems. A goal symbol is a symbol that can be created by the action of one or more productions, and whose presence in memory constitutes one of the conditions for the execution of other productions. When a goal symbol is placed in memory, the attention of the system is focused on those productions that include this goal among their conditions. It is only a short distance back from goals to traditional subroutine hierarchies; goals must be used with restraint if the "spirit" of decentralized control in production system architecture is to be retained.

In the SOAR system, described in Chapter 4 of this volume, productions execute in parallel. If this creates conflict -- an "impasse" -- then an explicit goal is created to resolve the impasse by problem solving within an appropriate problem space.

In symbolic networks, focus may be provided by specifying that conditions of productions can be matched only at nodes that are "active," and thereby serve as a momentary working memory. The procedures that cause the spreading and extinction of activation then become the main means of control.

In connexionist networks, the strengths of links and the levels of activity of nodes continue to change simultaneously until they match patterns that define the system goals and thereby reach equilibrium.

These brief remarks will provide some picture of the problem of control of cognitive activity and of the great variety of mechanisms that have been examined, and are being examined, to accomplish control and the focussing of activity. Much more detail about a number of specific architectures will be provided in the chapters that follow, especially Chapters 3 and 4, on connexionist architectures and SOAR, respectively.

Network Models

Only a bit more needs to be said here about network models, both connexionist and symbolic. But a bit of history may cast some perspective on them. The modern history of architectures for modeling cognition can conveniently begin with Donald Hebb's (1949) proposal of a "conceptual nervous sytem" in the form of a complex neural network. Higher mental processes were carried out in this network through the formation of cell assemblies: organized clusters of highly interconnected neurons. Some efforts (e.g., Rochester et al., 1956) were made to build computer models of a Hebbian system, but without notable success in showing that such a network could perform complex organized activities.

The network idea had, and continues to have, a number of attractions. First, however schematic the "neurons" or other elements of the network, it appears to hold some promise of bulding an eventual bridge to the physiology of real neurons. Most neurophysiologists, looking at existing network models, conclude that the bridge is still a long way off.

A second attraction of the network model is that it offers some mechanisms for learning: learning can take the form of appropriate changes in the activations of elements and the strengths of links. It is not surprising, therefore, that many of the tasks to which network models have been applied are concept learning tasks of one kind or another.

A third attraction of the network model is that it provides a direct means for modeling parallel processes like those that must go on in the eye and ear, and those parts of the brain closely associated with these senses. The Perceptron of

Frank Rosenblatt was an early class of connexionist devices aimed at modeling perceptual processes. Contemporary connexionist systems, though they incorporate a number of important new ideas and mechanisms, may be regarded as direct descendants of the Perceptron (influenced heavily, on the psychological side, by Hebb's ideas).

Symbolic networks can claim some of the same ancestry, but their development was also shaped by other influences. List processing techniques lend themselves, in the most direct way, to storing information in associative, semantic memories -- that capability was, in fact, one major motivation for their invention, and they have been used in this way from their beginnings in the 1950s.

Ross Quillian, in a Carnegie-Mellon doctoral thesis (1966), was perhaps the first to explore some of the psychological implications of this kind of memory model. His work initiated an active period of development of semantic memories of this kind, the introduction of the spreading activation mechanism (J.R. Anderson, 1983), and the application of these memories to the study of psychological phenomena of memory, language, and learning.

Symbolic Models

Again, since a good deal has already been said about symbol systems, we need only provide a few additional comments here on points not covered previously. Symbols are patterns with the special property that they designate or "point to" structures outside themselves. The structures designated by symbols may be other symbol structures in memory, but they may also be productions that can be used to recognize patterns in external stimuli (the objects denoted by the corresponding symbols). When symbols are organized into structures, arbitrary symbols (labels) assigned to these structures serve to designate or "name" them.

In most of the psychological models constructed as symbol systems, the elementary information processes have a "grain size" that corresponds to psychological processes requiring hundreds or a few thousands of milliseconds for their operation. In a few models (e.g., EPAM), some of the basic processes are at the 10-millisecond level, or thereabouts. Symbolic network models incorporate a substantial amount of parallel processing, but most other symbolic models are assumed to operate serially, with an architecture not too different from the classical von Neumann architecture.

In a serial system, it is an easy matter to model the limits of short-term memory and the effects of those limits upon cognitive activity. On the other hand, it is not as easy to model the massive parallel activity that goes on in the sensory organs, and perhaps elsewhere in the nervous system. A good deal of the discussion of the relative merits of symbolic and connexionist architectures centers on the question of the respective roles of serial and parallel processes in human cognition.

4. Two Approaches: "Reasoning" and "Search"

In the last section we discussed a variety of views in cognitive science on the architecture of intelligent systems. Cutting across this range of viewpoints lies another fissure that divides cognitive science and scientists into two groups who communicate with each other somewhat less than might be desirable. On the one side we have an approach that starts with language and logic and that views thinking as a process of inference or reasoning, usually employing a language-like representation. On the other side we have an approach that views thinking (especially problem solving and concept attainment) as a process of heuristic search for problem solutions, generally employing representations that model, in some sense, the problem situation.

It is not surprising that such a division should exist in a domain that has been assembled from the varied disciplines named in Section 2, or that the line of division should correspond, at least roughly to the disciplinary boundaries. For the most part, those who have come to cognitive science from philosophy, linguistics and psycholinguistics, and the more formal specialties within computer science find language and logic the congenial starting points for conceptualizing intelligence; while those who have come from artificial intelligence and most specialties within psychology are likely to employ the heuristic search paradigm as their point of departure.

For the moment, we will leave aside the third camp, the connexionists, who have ties with neurophysiology, and whose orientation, as we have seen, is computational, but whose representations are not symbolic.

The fissure between the logic/language and heuristic search viewpoints is somewhat visible in the chapter organization of the present volume, the chapters on semantics and logic, language acquisition, parsing, reading, and discourse mainly representing the language and reasoning approach, while the chapters on categories and induction and problem solving largely represent the heuristic search approach. The chapter on mental models lies somewhere between: Its very title associates it with heuristic search, but it clearly has its origins in the language-and-logic-oriented literature, and makes almost no reference to work employing models that came out of the other tradition in cognitive psychology and artificial intelligence.

In considering how these two viewpoints are likely to relate to each other in the future, several issues must be examined. The first is whether or not the human language faculty is a highly specialized and separate component of the human mind, in its semantic as well as its phonetic and syntactic aspects, operating with processes and on principles different from those implicated in non-linguistic thinking.

To the extent that language uses self-contained faculties that are independent of other parts of cognition, each paradigm might go its own way. To the extent that all modes of thinking, whether they involve language or not, share a common semantic representation and processes, there is a pressing need to find common ground where the two paradigms can meet.

A second issue, the central one, is to what extent "thinking" is to be identified with "reasoning," and "reasoning" with "logical inference." Is logical reasoning, in this sense, the correct metaphor for thinking, or is thinking better regarded as search? Others might prefer to state the question differently: are the "operators" that are applied in search to transform one situation into another to be regarded as "inference rules"? And what are the consequences of so regarding them? To some extent the competition between the LISP and PROLOG programming languages is a competition between viewing thinking as search and viewing it as reasoning.

Here, two questions are involved. First, does either one of these two paradigms give a uniquely correct picture of human thinking, or does thinking, on different occasions or in different aspects employ both modes, and perhaps others? Second, if there is both thinking in language and thinking without language (in images, say), how are these two kinds of thinking linked with each other?

A third issue, closely linked to the second one, concerns the representation of the knowledge on which thinking operates. Is knowledge in intelligent systems to be represented in the form of propositions or networks of propositions, on the one hand, or in the form of models, image-like or otherwise, of the problem situations, on the other. In particular, in human thinking, what are the forms of mental representation? Let us consider each of these three issues in turn.

Is Language Unique?

The hypothesis that language is a unique human faculty, and the further hypothesis that the capacity to use language is innate rather than learned, has been espoused strongly by Noam Chomsky (1965). At the outset, several points can be accepted as beyond debate. First, children obviously have to *learn* their native languages: whatever innate language capability they have is a capability for acquiring language and using it once acquired. And it appears that children learn any one of the several thousand native languages that exist about as easily as they learn any other. What are innate, if anything, are language universals and not the grammars of particular languages.

Second, parts of the human brain specialize in storing and processing the phonetic, syntactic, and lexical components of language, and these parts do not seem to be well represented in non-human mammals, even in the other primates. It is not a question of whether apes can or cannot acquire a modicum of language-using skill. Even if this has been demonstrated (it is still controversial), there is an enormous quantitative and qualitative difference between human language and the language that can be acquired by any other species. (Machine language raises other issues, which will be discussed briefly below.)

Given these two areas of agreement, what is still in contention is whether the innate language capacity is a wholly distinct faculty, or whether it is simply a gradual commitment, through learning, of part of the general human cognitive capacity to knowledge of language and language processing. The correct answer, of course, might fall somewhere between these two extremes.

On the sensory side, a human being must acquire the ability to discriminate among phenomes in his or her native language, thereby to recognize lexical items, and to parse the syntax of incoming acoustical signals so as to extract their meanings. This same human being must also have, or acquire, the ability to discriminate among the features of visually presented stimuli, to recognize familiar kinds of objects, and to detect and recognize the relations among objects in a visual scene. It seems parsimonious, even apart from the empirical evidence of brain localization, to postulate that there are specialized mechanisms for processing linguistic messages, on the one hand, and visual stimuli, on the other.

Accepting this specialization, the principal remaining question is how the semantic components of these two systems are related. What is the relation between the meaning that is extracted from the sentence, "There is a cat in this room," and the meaning that is extracted from seeing a cat in the room? The relation must be quite intimate, since a person who sees the cat will immediately (i.e., in a few hundred milliseconds -- see Chase and Clark, 1972) conclude that the sentence is true, and a person who hears the sentence may form some expectation of seeing a cat (unless the room is supplied with the kinds of crannies in which cats love to hide).

Here parsimony appears to favor communality. That is to say, if the mental representation of the meaning of the sentence were identical, or essentially so, with the mental representation of the visual stimulus, then it would be easy to use visual sensation to test the truth of the sentence or the sentence to create the expectation. If there is no such common semantic representation, then there must exist a process for translating semantic knowledge acquired through the visual route into semantic knowledge acquired from language, and vice versa. A regard for Ockham's Razor makes that alternative seem unattractive. We prefer the hypothesis that there is only one meaning -- one unified system of semantics shared by all subsystems that deal with meanings.

Another line of evidence that there is a single semantic memory comes from experiments that show that subjects store the meanings of sentences presented to them rather than storing the literal verbal strings. Thus they frequently report having seen a particular sentence when they have only seen a synonymous one, or sentences from which the one in question could be inferred (Bransford & Franks, 1971). In other experiments (Rosenberg & Simon, 1977), subjects have been unable to remember whether they saw a particular sentence or a simple drawing expressing the same meaning; while bilinguals are frequently unable to remember whether a sentence they saw was in French or English.

That a common semantics is feasible has already been demonstrated by a number of artificial intelligence systems. Early examples include the GRANIS system of L. Stephen Coles and the ZBIE system of Laurent Siklóssy, both built in the 1960's (Simon and Siklóssy, 1972, Chapters 5 and 6). In Coles' system, the computer is given an ambiguous sentence together with a picture of a situation to which the sentence is supposed to refer. If the sentence is found to be syntactically ambiguous, the semantic information in the picture is used to choose among the alternative parses.

Siklóssy's system proceeds in the reverse order. It starts with a structured description of a situation in what amounts to a picture language, and a verbal description of the pictured situation in a natural language. As a series of such pairs is presented to it, it slowly learns to form the natural language sentence appropriate to the situation, and is then able to construct such sentences when new pictures are presented without language.

A decade later, Hayes and Simon (1974) constructed the UNDERSTAND system, while Novak constructed a system called ISAAC, both of which also illustrate how semantic capabilities can be linked to natural language inputs. The UNDERSTAND program accepts natural-language descriptions of simple, puzzle-like problems, and constructs from them models of the problem situations and operators for making legal moves. ISAAC, given physics problems described in natural language, constructs a problem representation that allows an affiliated program to draw a picture of the problem situation on the terminal screen.

Three of these four programs (the exception is GRANIS) use essentially identical semantic representations: list structures in the form of networks built up of lists and description lists (property lists). These list structures may be interpreted as "pictures" because they represent objects as nodes (with descriptions of features) and link them with proximate objects. They may also be interpreted as propositions, since each component can be viewed as an assertion either that a particular object has a certain property, or that a certain relation holds between two objects. Essentially (although quantifiers are not represented in these particular programs), they correspond to representations in the first-order predicate calculus.

The conclusion we draw from psychological experiments in which subjects compare the meanings of sentences with pictures, and from artificial intelligence programs that handle both language and "diagrammatic" representations is that there is a common semantic system that contains meanings, whatever the source of information from which they are derived. Language does not have a separate and unique semantics.

Inference Rules versus Change Operators

We come now to the question of whether thinking is best regarded as reasoning, or as search, or as some combination of these. The reasoning viewpoint is closely associated with the notion that meanings are represented as propositions, while the search viewpoint is associated with the notion that meanings are stored as mental models of situations. It has been debated (Pylyshyn. 1973; Larkin & Simon, 1987) whether there is really any operational distinction between these two viewpoints

-- in particular whether it is possible to determine empirically whether mental semantics are propositional or diagrammatic.

The answer is surely affirmative. Even if two representations are informationally equivalent (contain exactly the same information), they may be quite different from a computational standpoint. Some information that is explicit in the one representation may be implicit in the other, hence retrievable only with the help of more or less extensive computation. Likewise, information that is implicit in both representations may be much more cheaply retrievable in the one than in the other. For example, it may be far easier to determine that two lines intersect by referring to a diagram than by trying to prove the existence of the intersection from the given facts, stated propositionally, together with the axioms of geometry (Larkin and Simon, 1987).

When human thinking is viewed as propositional, and when the analogy of logic is used to characterize human reasoning, the resulting models of thinking usually acquire certain traits borrowed from the domain of logic. Since logic is centrally concerned with rigor and verification, it generally employs only a small number of simple rules of inference -- the rule of modus ponens and of replacement of variables in the system of Whitehead and Russell, for example. All other "givens" are represented by axioms. When only a small set of inference rules is admitted, it is easy to check the validity of each step of a derivation, but usually at the price of making the derivations very long. Hence, the process of verification is facilitated, but the process of discovering derivations is made much more difficult than if the system is generously endowed with inference rules.

The reliance in logic upon a small number of "reliable" inference rules is not an inexorable requirement, but is a habit or custom carried down from a traditional orientation toward verification rather than discovery. Any number of inference rules, incorporating any number of assumptions, mathematical and empirical, can be introduced into a logic. However, when this is done freely, so that much of the knowledge of the task domain is expressed as inference rules rather than declarative assertions, the resulting system no longer looks to us like a logic, but instead, like a set of operators for carrying out heuristic search on a model. "Modeling" is neither more nor less logical than "reasoning." In modeling we proceduralize much of the knowledge that is represented declaratively in reasoning. Moreover, in modeling, we may (but need not) represent information in propositional form.

The liberal use of inference rules, especial rules incorporating semantic knowledge, blurs the distinction between the language/logic paradigm of thinking and the heuristic search paradigm. Whatever difference remains between them must arise from the particular knowledge representations they employ. That brings us to the third question: proposition-like versus image-like representations.

Propositions versus Images

Almost all of us believe that we can form mental pictures of situations described to us, or retain mental pictures of scenes we have actually seen. We are also all aware of the homunculus fallacy: a mental picture cannot be a photograph that is scanned by a homunculus in the brain. But there is also a homunculus fallacy associated with the idea of storing propositions in the brain: we must also exclude a homunculus who would extract the meanings from the propositions by *reading* them.

Both homonculus fallacies are avoidable in exactly the same way. A mental picture or a set of stored propositions are simply symbol structures of some kinds. The information processing system contains a variety of symbol-manipulating processes for extracting information from structures of these kinds. Thus, a processor of propositions may be able to extract the predicate from a simple proposition and then access in memory a symbol structure that contains information about the meaning of that predicate. Similarly, a processor of diagrams may be able to extract a feature (an intersection of lines, say) from a symbolic representation of a geometric form, and then access in memory a symbol structure that contains information about that kind of feature.

There are many ways in which propositions can be stored in memory. They can be stored as strings, similar to natural-language sentences. More commonly, nowadays, they are stored in the notation of the predicate calculus, or as components of node-link networks. Each mode of storage calls for a corresponding set of processes to examine symbol structures and extract information from them.

Similarly, there are many ways in which pictorial or diagrammatic images can be stored in memory. They can be stored as rectangular arrays (rasters) of points (pixels). They can also be stored as node-link structures, where spatial information is incorporated in the structure of the network, linking adjacent components so that the spatial relations can be detected. Again, each of these modes of storage calls for a corresponding set of processes to extract information from it.

To say that "mental propositions" and "mental pictures" can be handled in basically the same ways is not to say that there is no significant difference between representations. On the contrary, our earlier discussion of the computational equivalence or inequivalence of representations tells us that our ability to extract information efficiently (or at all) from a representation depends very much on what processes are available for manipulating it.

Hence, by studying what kinds of inferences people make easily, and what kinds thay make only with difficulty or not at all, we can draw conclusions about the kinds of representations and operators they are using. Kosslyn (1980), in particular, has used this strategy to argue that people use mental images extensively, and has suggested the forms in which these mental images may be stored -- arguing that both node-link structures and rasters are involved. His arguments seem cogent to us.

Summary: Reasoning and Search

Both the reasoning and search viewpoints toward intelligence are well represented in this volume. In summary of our discussion here, we need make only three points. First, while there are clearly some mental functions, and a physiological base for them, that are specialized for handling language, the weight of evidence suggests that, at the semantic level, there is a single common set of mechanisms used for all forms of thinking, and these are essential for translation between propositional and image-like inputs.

Second, the boundary between logical reasoning and heuristic search of models is a matter of degree rather than kind. Most human reasoning uses a liberal store of inference operators with substantial semantic content, hence resembles searching a model more than it resembles carrying out logical deductions.

Third, the knowledge on which inference rules operate may be stored in memory in a variety of ways, including propositional structures, but also including structures that are better viewed as mental images. Perhaps there are others as well.

5. Methods of Cognitive Research

Since good science consists of asking interesting questions that give hope of being answered, good methodology is essential to good science. Not only do methods influence the types of questions we ask, but they also determine what questions we can aspire to answer. Cognitive Science has at its disposal a remarkably wide and varied pool of methods, ranging from computer simulation to naturalistic observation, from recording the electrical impulses of the brain to performing linguistic analyses, from using non-monotonic logics to collecting reaction times or verbal protocols.

Rather than attempting to examine all these research techniques, we have chosen to explore a few of them in some detail, concentrating on protocol analysis and computer simulation. In addition, we discuss content analysis and meta-analysis as potentially useful methods not yet widely used in cognitive science. Content analysis serves as a generalization of protocol analysis, while meta-analysis provides insight into the methodology of computer simulation.

We will also have something to say at the end of this section about contributions of the methods of philosophy and logic to cognitive science. Since other authors in this book have addressed most of the other standard methods of cognitive science, we will not discuss those here. In particular Bower & Clapper examine experimental methods in psychology, while Sejnowski & Churchland review the methods of the brain sciences.

Protocol Analysis

Protocol analysis, the use of verbal reports as data, has become an increasingly important (but often misunderstood) technique for studying human intelligence. Its importance stems from the fact that it is one of the few methods in Cognitive Science that gathers data with sufficient temporal density to test models that account for behavior nearly second by second (but not millisecond by millisecond).

Although verbal reports date back to the introspectionists, the use of such reports *as data* should not be misconstrued as introspection. Introspection took subjects' verbalizations at face value, as constituting a valid theory of their own thought processes. Protocol analysis today, however, treats verbal reports as a source of data to be accounted for with an experimenter-generated theory -- perhaps in the form of a computer simulation.

There are many ways of getting subjects to generate verbal reports, including questioning them, asking them to report on their mental processes, and asking them to talk or think aloud. Verbal reports of the talk- or think-aloud variety tend to be the least intrusive and to provide the most accurate records of the nature and sequence of subjects' mental processes. We will focus our discussion on them.

Concurrent protocols can be obtained by asking subjects to think aloud *while* performing some task such as problem solving. Typically, the verbalizations are recorded on tape, transcribed, coded from the transcript, and then subjected to a theory-based analysis.

Sometimes, subjects are asked to report everything they can recall about the problem solving episode immediately *after* solving the problem. Such *retrospective* protocols can be useful for filling the gaps in concurrent protocols and in providing the subject's overall conception of the task. However, they must be interpreted cautiously, since subjects are quite capable of reconstructing events that did not actually occur (Nisbett & Wilson, 1977).

Two generalizations may be made about retrospective protocols. First, the longer a subject delays providing a report after completing a task, the more susceptible that report will be to reconstruction and distortion. Second retrospective protocols that might constitute questionable independent evidence can nonetheless be supported and confirmed with other data such as reaction time measures or error analyses.

Typical Procedures. In the paragraphs that follow, we discuss procedures typically used to study human information processes by protocol analysis (Ericsson and Simon, 1984). Depending on how the instructions are worded, the behavior of subjects may range from simple vocalization of verbally encoded thoughts as they occur, to elaborate introspections of the "I think I am using a perceptual process now ..." sort. The instructions typically discourage the the latter type of reporting and encourage the former. The exact wording of the instructions depends on the nature of the task and the degree of completeness desired, but the simple instruction to "talk aloud" while performing the task captures the essence.

After the instructions have been given, subjects are usually allowed to practice before actually performing the task of interest. If practice effects are a concern, a different task is used as a warm-up. One task that has been employed successfully is to ask subjects to take an imaginary walk through a familiar house or building, describing what they see. Another is to to have them multiply a three digit number by a two digit number mentally while talking aloud.

If subjects fall silent[2], the experimenter may remind them to keep talking. A non-directive prompt (e.g. "Keep talking") is less likely to interrupt the normal sequence of processing then a more directive prompt (e.g. "What are you thinking about?").

After transcription the protocol can be segmented and encoded. Segmentation involves dividing a protocol into units that can be encoded more or less independently. Criteria for segmentation depend upon the type of encoding and analysis. Typical cues for segmentation include completion of sentences or clauses, pauses, or completion of ideas.

Encoding is the process of translating the content of segments into a formal language. Determining what to encode depends largely upon the nature of the task and the hypotheses of interest. Choosing the level of analysis is the first step. To

[2] say, for 15 seconds to 1 minute, depending on the task

test a simulation of the subject's behavior, the protocol must be coded in sufficient detail to provide data at the level of the model. For example, a simulation that accounts for individual differences between subjects requires a more detailed encoding than a simulation that captures only behaviors common to all subjects.

Not all analyses are carried out at a fine level of detail, for one may be more interested in the overall sequence of goals and subgoals than in just how they are achieved. When aggregating the data from many subjects who have all solved a particular problem, one can simulate the general solution path first and then build more detailed variants of this general simulation to describe individual differences.

Verbal protocols generally provide explicit information about the knowledge and information heeded in solving a problem rather than about the processes used. Consequently, it is usually necessary to infer the processes from the verbal reports of information heeded instead of attempting to code processes directly.

The coding formalisms can be constructed from the actual vocabulary used by subjects, economising by defining equivalences between words used nearly synonymously. The particular words used provide powerful clues both about subjects' representations of the task and about their processing. For example, the tenses of verbs can help distinguish the planning of future actions from the retrieval of past information (e.g. the difference between "I will have" and "I had"). Function words (e.g. "if", "perhaps", or "so") can often distinguish exploring a hypothesis from drawing a conclusion, and so on.

The proper use of context in coding segments presents some perplexing problems. Inasmuch as segments sometimes refer to other segments, potentially important information may be lost if context is not taken into account. On the other hand,the number of independent observations of behavior represented by an encoded protocol decreases whenever segments are not coded independently of each other.

Potential loss of information must be balanced against potential gain in reliability when deciding just how much context to use in encoding.

Finally theoretical constraints often serve to focus attention on the relevant aspects of the protocol. For example, viewing problem solving as search through a problem space suggests coding instances of goal setting, and operator application. Similarly, if behavior is coded using the functional notation: R(x,y, ...), where R specifies a relation between two or more arguments, attention is thereby focused on the relations and their arguments.

Several strategies are used for determining the coding categories. First, they are often derived from a task analysis. An important advantage of this method is that then the coding categories are not developed from the protocol itself, hence are not ad hoc and do not subtract degrees of freedom from the data.

When it is not possible to develop coding categories solely from a task analysis, they are typically derived from the data that are to be coded (or some part of them). Since degrees of freedom are lost in this way, the goodness of fit of the theory should be judged by the number of residual degrees of freedom in the data (see Simon 1977 and Ericsson & Simon 1984 for a more complete argument). Making use of *simple* coding categories and procedures, it is quite possible to produce models that are parsimonious in relation to the amount of data explained, and hence quite credible.

Another observation concerns the derivation of categories from the protocols themselves. Using categories that are similar to the actual verbalizations both avoids loss of information and promotes reliablity in coding. Reliability in protocol analysis ensures that what otherwise might be criticized as idiosyncratic interpretation can be treated as "hard" data. Reliability is secured by using a simple and precise coding scheme.

Although inter-coder reliability in protocols is typically quite high (.8 or .9 is not uncommon), it can probably be increased by using automated and semi-automated coding systems. Several efforts have been made in this direction with some success. **PAS I** and **PAS II** (Waterman and Newell 1971, 1973) were designed as completely automatic programs for encoding segmented protocols and producing problem behavior graphs as output. These programs, while promising, require further development if they are to rival the performance of expert human coders.

SAPA (Bhaskar and Simon 1977) and PAW (Fisher 1988) represent a different approach to computer-aided encoding of protocols. Both of these programs interact with the human user, serving as powerful tools for protocol analysis rather than as fully automated systems. Fisher's system, for example, allows the user to devise the encoding scheme either in advance, from a task analysis, or while coding. The coding categories can be organized in schemas of a quite general kind, so that the system puts relatively few constraints on the encoding decisions. Procedures for analysing the encoded data qualitatively and quantitatively are incorporated in the system.

The Theory of Thinking Aloud. The procedures described above have been derived largely from experience. However, most of them can be induced from an information-processing theory of the processes employed in thinking aloud (Ericsson & Simon 1984).

The theory starts with the assumption that the processes that generate verbal reports are a subset of the processes that generate all observable behavior. and thus are amenable to an information processing analysis. It also assumes the existence of a Long Term Memory (LTM) and a Short Term Memory (STM). STM is limited, information enters it serially, and it takes time to attend to stimuli. The additional assumption that problem solving occurs in a problem space can be added to those mentioned above, but is not necessary for most purposes.

The key feature of this theory of verbal reports is that subjects can only report the contents of short term memory. There are three kinds of verbal reports. The first kind, *direct verbalization*, is the simple vocalization of information that is already encoded verbally in STM. Because the subject is vocalizing what is already in memory, such vocalization has no significant effect on the time to perform a task or on the sequence of processing. Subjects who talk to themselves naturally while solving problems are exhibiting direct verbalization. Similarly, subjects who are instructed simply to "talk aloud naturally" -- while avoiding introspection, reporting on their thoughts, or striving for completeness -- would be most likely to engage in direct verbalization.

The second kind of verbalization, *recoding the contents of STM*, includes verbal reports describing the contents of STM when these are not encoded verbally. Thus subjects instructed to think aloud while performing a task involving imagery may have to recode some of what is in memory before being able to give a verbal report. The theory states that the recoding processes will time-share with the processes involved in performing the task. In this case we would expect verbalization to slow down task performance, but not necessarily to change the sequence of processing. Because of the time-sharing assumption, the theory predicts that verbalization will cease under high processing loads that do not naturally produce verbally encoded information.

Experience with gathering protocols tends to validate this prediction. For example in the protocols of subjects trying to find a way to cover a "mutilated" checkerboard with dominos, considerable pauses often correspond to what appears to be imaginal or visual processing. When subjects are prompted in these cases to "keep talking," they often report that they are imagining different coverings or looking at the board and waiting for something to occur to them (Kaplan & Simon, in preparation).

The third kind of verbalization, *explanation*, includes verbal reports that require processing beyond simple recoding. For example, asking subjects to explain their thoughts may result in their trying to synthesize a coherent explanation. When additional processes are invoked as a result of the verbalization instructions, we expect the execution time of the task to increase. We also expect that the sequence of processes normally used to perform the task will be disrupted. As for the explanation itself, it may or may not be an accurate report of the processes it tries to describe. Much will depend on the task, the subject, and the nature of the attempted explanation.

Seemingly harmless instructions (e.g. "What are you thinking now?") can, by suggestion, launch a subject into a bout of explanation. While the explanations are often interesting in their own right, if our primary concern is to obtain a valid record of the sequence of information heeded (and processes used) in performing the task, verbalizations of the other two kinds are preferable.

A natural consequence of the assumption that verbal reports reflect the contents of STM is that subjects must be unable to report on processes that bypass STM (e.g. the cues they use in direct recognition), although they can report the outcomes of such processes if these are stored in STM. Since information must be attended to in order to enter STM, and since we assume seriality, the contents of STM also provides a sequential record of the information attended to.

With this theory in hand, it is a straightforward matter to derive several of our earlier assertions. The effects of instructions stem directly from triggering different kinds of verbalization by slight changes in wording. Warmup trials ensure that the subject has interpreted the instructions in the desired way and is verbalizing appropriately. Knowing that verbalization reflects the contents of STM explains why protocols provide better records of information attended to than of processes used. While many processes bypass STM, the results of these processes (i.e. new knowledge) are often deposited in STM. However, if one *is* interested in processing, the theory also helps interpret pauses and the processing deviations that can be expected with the different kinds of verbalization.

The theory helps to establish boundary conditions for the methodology. We have already noted that protocol analysis is not likely to be of use in examining processes that bypass STM, such as direct recognition. We have also shown that tasks involving imagery or non-verbal processing are likely to yield incomplete protocols with intervals of silence, since STM recoding processes will have to time-share with the processes used for performing the task.

Besides requiring recoding and/or the invocation of additional processes to form explanations, overt verbalization may affect what gets encoded in LTM or stays active in STM. The theory suggests that if only task relevant information is verbalized, verbalization could facilitate performance on a task by improving memory. Conversely verbalization of irrelevant information might interfere with the memory of task-relevant knowledge.

In using any methodology, we must ask to what degree data can be separated from theory. What is a reaction time without a theory relating time to amount of processing? What is an image from a microscope without the optical theory that explains what a microscope does? Both the collection of reaction times and the observations made with a microscope rest on theoretical assumptions that consume degrees of freedom. When conducting a protocol analysis of a new domain, it is particularly important to minimize the degrees of freedom consumed, since their loss must be balanced by the rather modest amount of data that the theory accounts for.

Fortunately, there are a number of ways to minimize the loss of information.

A first strategy is to keep theoretical commitments to a minimum, and to be explicit about assumptions. A second is to account for as much data as possible with a single set of assumptions. With regard to protocol analysis this means that the same coding procedures should be used over many segments within each protocol, and that protocols from different subjects should be coded with the same procedures. The same theory or variants of a more general theory should be used to account for data from different subjects on the same task. By keeping theories as similar as possible, one minimizes the number of degrees of freedom consumed. Converging evidence (e.g. reaction time measures, task analyses, error analyses, etc) can lend support to the theory. None of these principles, of course, are peculiar to protocol analysis. They apply to all methodologies.

While we have discussed protocol analysis in the context of typical psychological experiments (e.g. problem solving), this method is applicable in a wide range of domains. One brief example drawn from second language research will illustrate the possibilities. Gerloff (in Faerch & Kasper 1987) had subjects think aloud while translating a short French passage into English. Her results gave evidence for differences between good and poor translators with regard to the preferred language of analysis, the size of the units of analysis, and the characteristic patterns of movement through the text. Most interesting from a methodological perspective was her combined used of concurrent protocols with linguistic analysis in terms of morphemes, words, phrases, and so on.

Content Analysis

The strength of protocol analysis lies in providing a dense stream of data about human behavior over periods of minutes. This same principle of density of data can be used to study human behavior on even longer time scales, employing *content* *analyses* of articles, texts, and even primary documents such as laboratory notebooks. In fact, several historians of science now make systematic use of such logs in tracing the course of discoveries (Holmes, 1979); and in at least one case (Kulkarni and Simon, 1988), a computer simulation of a discovery process has been developed from, and tested against, a history extracted from laboratory logs.

In general, intellectual history would appear to be an excellent domain for increased interaction between cognitive scientists and historians. Advances in our understanding of cognitive processes should provide new ways for analysing the differences in modes of thought of different cultures and different eras. Conversely, historical texts can provide voluminous materials for testing hypotheses on how people represent and solve problems. However, cognitive scientists have yet to make full use of the methodologies that might aid in research of this type. Content analysis is one such option that has been largely overlooked.

Stone defines content analysis as "... any research technique for making inferences by systematically and objectively identifying specified characteristics within text" (Stone et al. 1966). In practice, content analysis proceeds very much like protocol analysis, but in the domain of written text.

Weber (1985) outlines some typical steps in performing a content analysis. First one defines the size of the unit one wishes to code (e.g. word, meaning of word, sentence, paragraph, etc.). Next one defines categories for coding these units. Now one is ready to test the coding scheme on a small sample of text. After the pilot test, one assesses reliability of the coding scheme and revises the procedures if necessary. After a number of loops through the revise-and-test cycle the text is ready for final coding. The last step is to assess the reliability actually achieved.

As in our earlier discussion of protocol analysis, the procedures just outlined raise a number of issues. In choosing the size of the unit to code we face the same context vs. reliability tradeoff. As Weber (1985) states, "large portions of text such as paragraphs and whole texts are usually more difficult to code as a unit than smaller portions such as words and phrases, because large units typically contain more information and a greater diversity of topics." On the other hand, single words or phrases often do not include enough context to disambiguate meanings. In analyzing protocols, sentences or complete ideas are often considered as segments. These criteria would seem appropriate for many content analyses as well. However, as Weber emphasizes in his treatment of the subject, the resolution of such issues in each case must depend primarily on the goals of the research.

Another issue involves the scope of the categories. If the categories are too wide, one loses resolution, but if they are too narrow, the power of abstraction is lost in the proliferation of categories.

There is also the issue of the exclusivity of categories. Should a unit of information be encoded into a single category, or if it seems to fit more than one category should it be encoded in all of them? One solution is to encode in multiple categories but to assign weights to the units based on how well they fit the categories. Other options include hierarchical coding schemes involving multiple classification in first and second order categories. The general disadvantages of decreased reliability and "fuzzy" categories must be set against the benefit of completeness.

Reliability is a central issue in both protocol analysis and content analysis. Weber refers to three types of reliability in content analysis: stability, reproduceability, and accuracy. Stability is essentially test-retest reliability, that is, the degree to which a given coder will code items the same way on two different occasions. A more stringent test of reliability is reproduceability or inter-coder reliability: the degree to which two or more coders agree on the way items should be coded. Finally, accuracy is the degree to which human coder performance matches some established normative encoding for the items. Since most studies involve coding new information rather than coding information for which normative encodings have been established, accuracy is not often a useful test of reliability.

As with protocol analysis, computer programs offer tremendous opportunites for improving reliability in content analyses, providing, in fact, a way of achieving essentially perfect coding reliability. (Validity, of course, may be more problematic.) Programs like the General Inquirer (Stone et al. 1966) greatly facilitate the tedious process of identifying and tabulating instances of categories in text for such domains as political science, personality psychology, social psychology, clinical psychology, and anthropology. With the development of relatively cheap optical scanning procedures and artifical intelligence programs that can actually draw inferences from text (see Dyer 1983), the potential for computer aided content analysis is enormous.

For example, to deal with the conflict between coding in unique or multiple categories, the computer can efficiently code the data both ways, allowing the results to be compared. Similarly, the issue of context becomes less of a problem if one has the option of retrieving additional context and recoding at the touch of a button. Because these are issues of theoretical as well as practical import, they cannot be settled automatically. Nevertheless, AI techniques have taken much of the drudgery out of content analysis, leaving a methodology with great promise for cognitive science.[3]

AI & Computer Simulation

Artificial intelligence is concerned with programming computers to perform in

[3] For more details on the uses of content analysis we refer the reader to the excellent reviews by Holsti (in Lindzey & Aronson ed. 1968).

ways that, if observed in human beings, would be regarded as intelligent (Simon in Corsini 1987). Adding the qualification that the computer programs ought to succeed and fail in exactly those places that humans do, we can extend this definition to cover computer simulation of human behavior as well.

If we want to understand the relation between the techniques of AI and modelling of human havior, we can compare a "pure" AI program with human behavior. Paige and Simon (1966) did exactly this in their investigation of the cognitive processes used in solving algebra word problems. Starting with STUDENT, an AI program developed by Bobrow (1963) which solved algebra word problems based on purely syntactic analysis, Paige and Simon compared STUDENT's performance with verbal protocol data collected from human subjects. They found that while some subjects solved the word problems in much the same way as STUDENT, others used auxiliary cues and physical representations as an aid to their solutions. Besides investigating the link between protocol data and computer modelling, Paige and Simon showed that a program imported from artificial intelligence could make testable predictions about human behavior.

The overall method of computer simulation is well illustrated by a more recent computer simulation. READER (Just & Carpenter 1987) is a production system model of some of the processes involved in reading and understanding text. In particular, READER simulates the time course of human reading, taking more time to process difficult words just as human readers do. READER also builds a representation of the text as it reads, with the result that it can answer text-related questions and produce summaries similar to those produced by human subjects.

The model incorporates processes that encode words, access a lexicon, perform semantic & syntactic analysis, utilize schemas, and construct a referential representation of the world described in the text. The number of cycles of production firings in READER reflects the amount of processing being done by the model. Since eye movement studies indicate that gaze duration on a given word reflects the time needed to perceive, understand, and integrate that word, it is a relatively simple matter to compare the number of processing cycles spent by READER with the amount of time spent by human subjects on a sample passage of text. Just and Carpenter made this word-by-word comparison on a 140 word passage about flywheels. They report that READER accounts for 67% of the variance in the human (mean gaze duration) data.

The strengths of the model do not rest upon fit to the data alone. Rather READER's ability to simulate and coordinate several different processes in parallel must also be taken into account. While it may be comparatively easy to simulate any single process, building a plausible *integrated* system that accounts for human behavior is a much more difficult task.

Finally, Just and Carpenter make claims for the generality of the production system architecture underlying READER, named CAPS (Collaborative Activation-based Production System). In particular they point out that CAPS has been used to simulate other types of processing (e.g. mental rotation tasks) besides those addressed by READER. Their approach reflects a growing trend in computer simulation to strive for generality of an architecture across domains (see Newell's Chapter 4 for a more detailed example).

READER is instructive in typifying a number of stages in the process of building a simulation. In the case of READER, Just and Carpenter had already developed a fairly well defined theory of the processes used in reading (Just and Carpenter 1980). READER sought to simulate reading in such a way as to predict gaze-duration data which their theory explained. In building it, both theoretical assumptions and constraints of the architecture were taken into account (Thibadeau, Just & Carpenter 1982).

If READER had been simply tuned to the specific performance data, rather than incorporating more general theoretical assumptions (e.g the immediacy of processing) then we would be less impressed by its fit to the data, for tuning consumes degrees of freedom. If models have free parameters without any theoretical means of determing their values, tuning is unavoidable, for the parameters must be set empirically. In these cases, the problem can be ameliorated if the tuning uses only a *portion* of the available data and is then tested on the remaining data.

After constructing the simulation, the performance of the model was compared with data from human experiments. In addition to quantifying the goodness of fit, it is important to pay attention to discrepancies between the model and the human data, as well as to predictions of the model that cannot be tested with the current data. Discrepancies indicate how the model might be improved; while predictions can suggest and guide further experimental research.

One way in which READER's encoding of words differed from that of human subjects was that READER encoded almost every word. Human subjects, however, skip about 60% of the short function words. One explanation for the difference is that human subjects may encode short words in parafoveal vision while fixating on the preceding word. The discrepancy between READER and the human data produced a version of READER that can use peripheral vision in skimming a text (Thibadeau, Just, & Carpenter 1982). This new research illustrates how different versions of a model, each incorporating different theoretical assumptions, can often provide more insight than a single model.

If provided with new tasks, a simulation has the potential for responding in new ways which may then suggest experiments with human subjects to check the predictions. Trusting the model transforms computer simulation from a tool that helps refine a theory to a source of prediction that drives research. Full exploitation of computer simulation requires viewing a model not as an end product, but rather as a continually evolving entity whose permutations may suggest new investigations.

The final stage in simulation is also the first. That is, once a model has been built and tested, it is time to build a revised version or a different model altogether -- either to generalize the model and architecture by applying it (or variants of it) to related domains, or to test its new predictions.

Many theoretical flaws can remain hidden in a verbal theory until one begins to program it. Then it becomes apparent that not any old program will fit a lengthy protocol, but that it is usually extremely difficult to build even one simulation that can account accurately for the data.

While we have discussed what we consider to be a particularly fine example of computer simulation, not all the models being published today live up to its standards of predictive power, generality and ability to account for human data. The techniques of meta-analysis, to which we turn next, provide a way of examining the characteristics and qualities of simulation models and experiments that are to be found in the literature.

Meta-Analysis

Meta-analysis, or more generally, quantitative reviewing, attempts to bring the precision and reliability of experimental methodology to the task of reviewing and synthesizing groups of studies. The basic procedure in meta-analysis is to define a set of questions that the analysis will address and develop a scheme for coding each relevant study in the literature. Statistical methods exist for combining the results of several different experiments addressing the same issue while taking the experimental designs, the magnitudes of effects, and other factors into account (see Wolf 1987 for a concise review). However, the power of meta-analysis lies not only in statistical comparision but also in coding schemes that permit quantitative answers to questions about groups of studies.

As a demonstration of the applicability of the method, we conducted a small scale meta-analysis to answer the questions raised in our discussion of computer modeling above (For a more complete report see Kaplan 1987).

Since we were interested in computer simulations as psychological theories, we decided to limit our meta-analysis to simulations that refer to actual human data (although they did not always compare model with data). In keeping with the scope of our ambitions (a demonstration rather than a full-fledged study) we considered simulations from a single journal only, *Cognitive Science*, and from a limited time period, 1980-1986 inclusive. While *Cognitive Science* is probably the richest single source of computer simulations of human cognitive behavior, we recognize that our sample is quite narrow, and only intend that the analysis provide some "ballpark" figures indicative of general trends in the use of computer simulations.

We categorized simulations according to the type of architecture used to build the model, the type of human behavior modeled, the time scale of the behavior modeled, and the degree of predictive power, generality, and fit to the data claimed by the authors.

Twenty three articles that met the criteria for inclusion were coded along the dimensions mentioned above. Table 1 shows the number of simulations falling into various categories. From the data in the table we reached the following conclusions:

First, the majority of simulations that had actually been built were either completely tuned to the data or compared to existing data with no mention of testable predictions that might be followed up. While some of simulations appeared to meet high standards of predictive power as well as of generality or of goodness of fit to the data, it was rare to find a simulation meeting all of the normative criteria discussed above.

Second, claiming a qualitative fit to the human data seemed to be the norm. Most of the authors presented neither numerical predictions nor quantitative comparisons of their model's performance with human data, yet still claimed that the model performed the task in a psychologically plausible way. There were approximately twice as many simulations claiming qualitative fits as there were simulations claiming quantitative fits. When statistics were provided comparing the models' behavior with that of human subjects, the models tended to account for between 50% and 80% of the variance.

Table 1

Distribution of Simulations among Various Categories

CATEGORY	FREQ. BY DIFFERENT TYPES OF SIMULATION ARCHITECTURES			
	# SERIAL	#PARALLEL	#OTHER*	TOTAL
Predictive Power:				
not actually built	3	3	0	6 (26%)
tuned to the data	6	1	0	7 (30%)
compared w/o tuning	4	3	1	8 (35%)
predictions made	1	0	1	2 (9%)
Good. of Fit to Data:				
no comp. w/hum. data	5	1	0	6 (26%)
qual. fit claimed	5	5	1	11 (49%)
quant. fit claimed	4	1	1	6 (26%)
Generality of Model:				
no discussion of	1	2	1	4 (17%)
general discussion of	9	3	0	12 (52%)
claims for	1	1	1	3 (13%)
demonstrations of	3	1	0	3 (17%)
Generality of Archit.:				
no discussion of	6	5	1	12 (52%)
general discussion of	7	2	1	10 (43%)
claims for	1	0	0	1 (4%)
demonstrations of	0	0	0	0 (0%)
Domain Simulated:				
recognition	0	1	0	1 (4%)
skill acquisition	1	0	0	1 (4%)
skill performance	0	1	0	1 (4%)
reading	0	0	1	1 (4%)
language generation	0	2	0	2 (9%)
imagery	0	1	1	2 (9%)
perception	1	1	0	2 (9%)
memory	3	0	0	3 (13%)
language comprehen.	3	1	0	4 (17%)
problem solving	6	0	0	6 (26%)
Time Scale of Behav.:				
O[100 msec.]	0	3	0	3 (13%)
O[1 sec.]	2	4	1	7 (30%)
O[10 sec.]	3	0	0	3 (13%)
O[minute(s)]	8	0	1	9 (39%)
O[hour(s)]	1	0	0	1 (4%)

Total serial simulations = 14; Total parallel simulations = 7
*Other category consists of one serial/parallel hybrid, and one model whose architecture was not described in sufficient detail to make the serial/parallel judgement.

Third, the generality of the models were usually discussed only in broad terms, with only 17% of the models actually demonstrating generality within or across domains. As to the generality of architectures, about half of the authors did not discuss this matter, while the other half talked mainly in general terms.

Fourth, simulations have covered a wide range of human behavior, with concentration of effort on problem solving and language comprehension, and on behavior that is executable in a matter of minutes or behavior transpiring in about 1 second.

Finally, while parallel simulations covered a wide range of human behavior (problem solving and memory being notable exceptions), they modeled exclusively phenomena occuring on the time scale of 100 msec. to 1 sec. Conversely, serial models simulated only behavior taking a second or more. The sharp division between parallel and serial models with respect to the time scale of the behavior modeled serves both to confirm intuitions about the appropriateness of these architectures for tasks at different grain sizes, and to highlight the challenge that any "grand scale" serial or parallel architecture must face -- breaking the time scale barrier. Serial models scored slightly better than parallel models on generality and goodness of fit, although the sample size is hardly large enough to draw firm conclusions.

In our meta-analysis, two generalizations appear noteworthy. First, a large proportion of the computer models in our sample lack the generality and predictive power that we would hope for. The remainng minority, however, illustrate what we ought to aim for in building simulations of human data. Second, there is a striking division between the time scales of phenomena modelled by parallel and serial models, respectively. This division supports the levels argument made earlier in this chapter.

We hasten to qualify these conclusions however, because of the limited scope of our demonstration. With regard to generality our estimates may be low, since authors might subsequently publish extensions or variants of their models in other jounals, or may omit references to simulations published elsewhere. Further, some simulations claiming only a qualitative fit to the data, but modeling several processes, might have to meet more stringent criteria than simulations that fit less demanding data quantitatively. Although it is our impression that such factors did not play a large role in the 23 studies we considered, issues like these should be resolved before strong conclusions are drawn from the analysis.

Our main purpose has been to demonstrate that a meta-analytic review allows us to answer questions both more objectively and more precisely than a standard literature review of comparable scope. Quantitative reviewing techniques like meta-analysis (See Green and Hall, 1984, for others as well) are likely to play an increasingly important role as cognitive scientists seek to integrate knowledge from the various literatures that are their heritage.

Philosophy, Logic, & Semantics

Historically, philosophy has been wholly eclectic in its methods. Philosophers appeal to everyday observation and introspection. They employ dialectical methods and seek out ambiguities and common confusions. In their search for more rigorous methods of reasoning, they have contributed to cognitive science (as well as to other disciplines) the powerful tool of modern symbolic logic.

However, with the very important exception of symbolic logic, philosophical methods of inquiry probably have less impact on the methods of cognitive science, than do the substantive questions that philosophy addresses: for example, the age-old questions of the nature of mind and intelligence, of the relation of mind to body, and of how we come to know the external world.

The qualities (if any) that make man unique in the universe have long been of interest to philosophers. Since the abilities to think and to use language intelligently have been regarded as central to the claims of human uniqueness, some philosophers have sought to establish boundaries beyond which the intelligence of machines cannot venture (Searle, XXXX, and Dreyfus, 1972). Although one need not agree with their claims about the limits of machines (as indeed we do not), the debate has been of some value in setting tasks for artificial intelligence research. If the claim is that machines cannot do X, then a computer program to do X becomes an interesting design and construction problem. One could write a good deal of the history of AI and contemporary cognitive science in terms of responses to this challenge (Simon, 1987).

The debate on limits has also been of some value in forcing clear operational definitions of such terms as "thinking," "understanding," "intuition," and the like -- definitions that allow unequivocal decisions to be made as to whether a particular computer program is exhibiting one or another of these qualities (or whether a human being is, for that matter). In the absence of such definitions, of course nothing can be settled.

During the brief history of AI and cognitive science, one after another of the initial philosophical distinctions between human and machine intelligence have fallen by the wayside as the technology has advanced (Simon, 1979b). Thus even as philosophy contributes to cognitive science by debating the limits of machine intelligence, cognitive science has strongly affected philosophy by producing running computer programs whose performance bears upon classical philosophical issues.

Additional evidence of the relevance of cognitive science to philosophy can be found in Section 4 of this chapter, where we discuss such topics as the relation of language to other cognitive faculties, reasoning and search as models of thinking, and thinking in propositions versus thinking in images. As Pylyshyn and XXX indicate in their chapters in this book, all of these questions are of great concern for epistemology.

Returning now to the methods of philosophy, we observe that modern symbolic logic has been absolutely fundamental in providing foundations for computer science in general and AI in particular. Apart from the use of the formalisms of logic in programming (e.g., the contribution of the lambda calculus to the LISP programming language, or the role of the first-order predicate calculus in PROLOG), the two key contributions have been Gödel's incompleteness theorems and Turing's (and Church's) work on computability.

The Gödel theorems have often been the basis for arguments on the limits of computers. These results show that in rich systems of logic certain theorems, known to be true by reasoning in a metalanguage, cannot be proved within the logic itself. What is sometimes overlooked is that the Gödel theorems are wholly neutral as between people and computers, placing just as strict, but no more strict, limits on the one as on the other. Hence the implications of the incompleteness theorems for the question of machine intelligence are still quite uncertain. The Turing machine provided a wholly new standard of precision for our examination of symbolic processing and its relation to thinking. But neither the Turing machine nor the other resources provided by symbolic logic have yet resolved all of these difficult questions.

The problem of reasoning under conditions of uncertainty is of interest to both psychologists attempting to understand human decision making and AI researchers trying to design systems that perform intelligently under real life conditions. We alluded earlier to the development in economics and statistical decision theory of an elaborate formal normative theory of decision making. In recent years, a good deal of psychological research has been carried out (Kahneman, Slovic & Tversky, 1982) that shows the limitations of this normative theory as a description of actual human choice under conditions of uncertainty. Some philosophers and some psychologists have participated in these developments, but on the whole, both the normative and the descriptive theories have been rather peripheral to the cognitive science endeavor.

In Section 4, we discussed logic as a possible model for the thought process -- with somewhat negative conclusions. We did not allude there to efforts, both in philosophy and in AI, to develop non-standard logics that would capture more of the properties of everyday reasoning than are captured by the predicate calculus in its usual forms. Modal logics, dealing with concepts like causality and possibility, have been created and extensively investigated (e.g., McCarthy and Hayes, 1969), but no consensus has been reached about a form of modal logic that would be especially helpful for or appropriate to the study of intelligence.

Work has also been carried forward on non-monotonic logics, designed to deal with the fact that large data bases, built up sequentially and constantly modified, may contain numerous contradictions; and that it may be advantageous to reason tentatively, preserving the ability to change one's conclusions when contradictions surface. Nonmonotonic logics have not yet found wide application, but they represent an interesting area of intersection of philosophy with AI.

Themes

Before leaving our discussion of methodology, we note several themes that permeate our discussion. One such theme is the principle of density of data. Protocol analysis provides a very large number of closely spaced observations on a single task. Similarly, content analysis and meta-anlaysis find high density of data in written materials. Content analysis extracts many data points from a single text, while meta-analysis considers each published result as a data point and examines many such results. Computer simulation requires a dense stream of data in order to constrain the processes that are being modeled. Whether one is interested in a single instance of human information processing or the broader picture composed of many such instances, one seeks methodologies that provide a dense set of data points at the level of interest.

A second theme is that boundary conditions for a methodology must be established before the full potential of the method can be realized. Protocol analysis illustrates this point most clearly. Reactions against earlier introspectionist methods contained some grains of truth, without a clear way of separating the wheat from the chaff. Hence, many psychologists decided to avoid the methodology altogether, or worse yet to condemn it. Our theory of concurrent protocols enables us to distinguish introspection from verbalization of the encoded contents of STM. The latter is likely to provide an accurate account of the information heeded during the performance of a task, while the former introduces all kinds of disruptive auxiliary processes which muddy the picture considerably.

The point is a general one. We are able to apply a method effectively to the extent that we understand its capacities and limits. The theory of protocol analysis exemplifies one approach to getting that understanding. Extending the concepts of protocol analysis to new domains (e.g. linguistics) or to new levels of human behavior (e.g. content analysis) is also a fruitful approach. One finds limits by pushing them. Meta-analysis can help by providing a good idea of where to start pushing.

The last theme is one of interaction and combination. Repeatedly in our discussion we have seen how methodologies from different areas have combined to create powerful new approaches. Viewing content analysis and protocol analysis as variants of the same underlying concept suggest that each might have ways of dealing with common issues (e.g reliability of coding) that the other could find useful.

Applying AI to content analysis and protocol analysis has produced semi-automated encoding schemes whose potential has just begun to be tapped. Finally, performing meta-analyses on computer simulation (or any methodology) helps provide the clear picture of a methodology that is prerequisite for fruitful interactions between the disciplines of cognitve science.

6. Invariance of the Laws of Cognition[4]

Intelligent behavior is adaptive, hence must take on strikingly different forms in different environments. Intelligent systems are ground between the nether millstone of their physiology or hardware, which sets inner limits on their adaptation, and the upper millstone of a complex environment, which places demands on them for change. Because they change in adapting to their environments, intelligent systems may be described as "artificial."

The task of empirical science is to discover and verify invariants in the phenomena under study. The artificiality of information processing systems creates a problem in defining such invariants. Any accurate statement of the laws of such a system must take into account their relativity to environmental features. We often find that we are studying sociology -- the effects of their past histories on our subjects -- when we are seeking to study physiology -- the effects of the structure of the human nervous system.

Finding invariants, then, in artificial phenomena is not easy. But the example of biology shows that it is not impossible. For biological systems were also absent at the creation of the universe. They too are shaped to their environments by the forces of evolution. Biological invariants (like psychological ones) hold for limited

[4] This section draws heavily upon Simon, (1980).

intervals of time and bounded regions of space.

On the shortest time scale, intelligent (adaptive) systems change their behavior in the course of solving each problem they encounter. A prime characteristic of effective heuristic search is that it shapes itself to the environment in which it finds itself. It tries to capture in its problem space the important features of that environment -- what Allen Newell has called the knowledge level.

On a longer time scale, intelligent systems make adaptations that are preserved and remain available for meeting new situations. They learn. There are many forms of learning. One important form is the accumulation of information in memories, and the acquisition of access routes for retrieving it. Learning changes systems semi-permanently, hence increases the difficulty of searching out invariants.

On the longest time scale, intelligent systems evolve: biologically by mutation and natural selection; and socially, through the accumulation and transmission of new knowledge and strategies. These changes further narrow the domain of invariance.

What room, then, is left for a general science of cognition? We must seek for invariants in the inner and outer environments that bound the adaptive processes. We must look for basic characteristics that might be held in common among diverse kinds of intelligent systems, and we must look for common elements at the knowledge level, among complex problem environments.

The Outer Environment

The first source of invariance in intelligent systems derives from common characteristics of the environments to which they must adapt. and in which their behavior takes place. Again, since these environments are highly diverse, the invariants will be correspondingly abstract. The environments that demand the exercise of intelligence are problem environments that do not present obvious paths to the attainment of a system's goals. Frequently, though not always, problem environments contain large, sometimes immense, numbers of alternatives, only a small fraction of which satisfy the goal requirements.

The principal mechanism of intelligence observed in people and computers operating in problem environments is heuristic search. The heuristic devices make search feasible by enabling the intelligent system to exercise great selectivity: (1) by using information stored in memory to choose more promising over less promising paths, and (2) by extracting from the problem environment new information about regularities in its structure that can similarly guide the search.

The interface between intelligent systems and their external environments -- their sensory and motor organs -- presents most delicate problems in the design of adaptive systems. For the interface must meet the requirements of both the outer and the inner environments -- must, in fact, communicate between them. It is no accident that our progress has been much slower in AI in designing effective sensors and effectors, or imitating the corresponding human organs, than in understanding and imitating those intelligent processes that can go on inside the human head without interaction with its environment. "Deep thinking" has proved easier to understand and simulate than hand-eye coordination.

The Inner Environment

We should not expect the invariants of the inner environment to be the same for all intelligent systems, yet there are one or two surprisingly general constraints shared by most of them.

For most purposes, we are concerned with just two kinds of intelligent systems: living organisms and computers. All living organisms make use of essentially the same protoplasmic material, but exhibit a wide variety of organizations. Most of the computers that have been built in the past forty years exhibit remarkable similarity of organization, but have been assembled from a most diverse set of alternative component materials. Hence, over these two classes of intelligent systems, we do encounter a considerable diversity of both organization and material subtrate.

Computers and (in our view) human beings are symbol systems. They achieve their intelligence by symbolizing external and internal situations and events, and by manipulating those symbols. They all employ about the same symbol-manipulating processes. Perhaps that particular invariance arose because computers were made in the image of man; but since only the connexionist view proposes a radically different architecture, perhaps the invariance goes deeper than imitation.

When required to do complex tasks, both people and computers exhibit vividly a severe attention bottleneck, or seriality. We may conjecture that the reason for the predominance of seriality at the level of complex processes (processes taking perhaps a half second and more in the case of human beings) is that it is very difficult to organize parallel computational systems if precise coordination is required of the computations being made simultaneously by the different components.

Where processing is basically serial, all of the relatively labile inputs and outputs of the basic processes can be handled in a working memory of limited size. (See the discussion, in Section 3 of this chapter, of short-term memory in the "standard model.") The need for a tradeoff between flexible adaptation to the environment and coherent attention to goals also seems to point toward mechanisms for attentional focus. Hence, when intelligence is implemented by production systems, a portion of memory is generally designated as working memory or as the "activated" portion of memory in which the condition sides of the productions must be tested.

In this limitation of the size of short-term memory, and in the consequent seriality in behavior, we see one of the invariants we may seek to characterize and understand. As can be seen from this example, the invariant is of a relatively abstract kind, imposing constraints upon the possible organizations of intelligent systems rather than upon their material substrates. Intelligent systems, according to this argument, will necessarily be symbol systems, and at the level of goal-oriented activity, they will be serial in operation with, narrow attentional focus.

Empirically Observed Invariants

However adaptive are the intelligent systems with which we are concerned, we should not despair completely of finding invariants in their behavior. For example, the study of expertise in humans and research on the design of expert systems for computers have revealed some remarkable uniformities and generalizations. These uniformities stem from both the limitations imposed by inner environments and the task demands of the outer environments.

Expertise, empirical research shows, generally rests on an extensive knowledge base. It has been shown for a number of fields of human endeavor that a world-class level of expertise is never attained with less than ten years of concentrated learning and practice. Here, the magic number 10 represents a ratio between the amount of knowledge and skill that has to be acquired by the expert (the demand of the external environment) and the rate at which people can acquire knowledge and skill (the limits of the inner environment).

With respect to the size of the knowledge base, it has been found that thousands or tens of thousands of productions must be provided to a sophisticated expert system like a medical diagnosis system. Research on human world-class experts has shown that their knowledge bases are likely to amount to 50,000 or 100,000 productions or more. Numbers of the same magnitude characterize typical natural-language vocabularies of professionals. The ubiquitous presence of large knowledge bases like these is a characteristic invariant of human and mechanical expert systems.

Possessing large knowledge bases of this kind, which operate very much in the manner of production systems, experts solve a great many problems by what can equally well be called "recognition" or "intuition." Even when problem solution requires extensive heuristic search, recognition plays a major role in allowing large steps to be taken "instantaneously," and permitting appropriate operators to be recognized and applied at each stage. Most of the AI expert systems have similar characteristics -- generally depending more on ability to recognize extensive sets of cues than on the ability to model problem situations at a deep level. Powerful recognition capabilities are another common property of expert systems.

We encounter numerous invariants, too, when we get close to the sensory and motor organs that form the interface between the inner and outer environments. Perhaps one reason why many of the invariants we have discovered belong to these interfaces is that they are more accessible to study than are the deeper regions of the brain. Another possible reason is that they are more specialized, through a long period of evolution, hence have more specific physiologically determined features than the more central and newer parts of the brain. Whatever the reasons, the relevant chapters of this volume testify that our knowledge of these interfaces is more precise than most of our knowledge of other aspects of cognition.

In the matter of machine intelligence, the reverse is true. The sophistication of the mammalian sensory and motor capabilities have proved exceedingly difficult to match in artificial intelligence programs. Hence we go about the design of "white collar" expert systems with far more assurance and success than we show in the design of robots.

Learning

The idea that intelligent systems are highly adaptive and flexible in their behavior could well lead to the notion that a great many of their invariants are to be found, not in their behavior when confronted with their usual tasks, nor in the structures responsible for performance, but in the long-range mechanisms that bring about adaptation -- their learning mechanisms. As a matter of fact, learning was a very popular topic in the early history of artificial intelligence research, and an almost dominating topic in experimental psychology during the period from World War I to the middle 1950s. The historical reasons for this domination are complex, and only one aspect needs comment here.

With respect to artificial intelligence, many early workers held the view that it was easier to induce a system to organize itself from scratch, by exposing it to an appropriate sequence of training experience, than it was to provide it with the knowledge it would need for expert performance. It was probably also thought that learner programs avoided the charge, commonly directed toward AI systems, that the intelligence was really in the programmer and not in the system. If the programmer only provided a potential for learning, and not the program for the finished performance, this accusation could not stand.

Whatever the reason, many investigators in the early years followed the learning route, but not with notable success. All of the expert systems that have been produced up to the present time (except for Samuel's (1959) checker-playing program) have been given directly all or most of the knowledge and strategy on which their expertness rests.

Within the past decade there has been a revival of the learning enterprise, but with viewpoints quite different from the original simple reinforcement paradigms. Connexionist architectures bring in a number of new powerful learning techniques -- for example, the so-called "annealing" or "Boltzmann" methods for causing a system to settle down to an appropriate equilibrium.

Within the domain of symbolic schemes, the new generation of learning programs, often using the production system architecture, are basically problem-solving systems, capable of undertaking heuristic search with a reasonable armory of methods like generate-and-test and means-ends analysis. They do not start out in the barebones fashion of the earlier generation of self-organizing systems. They also start out with a better picture of the goal of the learning: of the expert performance that the learner is striving to attain. Finally, today we have clearer ideas of the mechanisms (especially adaptive production systems) needed to enable a symbolic system to bootstrap itself.

We should not suppose that it will be simple to find invariants even in learning systems. After all, we must be prepared for the phenomenon of "learning to learn." The adaptive mechanisms themselves may learn from experience. In our study of performance and learning, we must not imagine invariants where there are none. Since intelligent systems are programmable, we must expect to find that different systems will use quite different strategies to perform the same task.

Hence, research on adaptive systems must take on a taxonomic, and even a sociological, aspect. We have a great deal to learn about the variety of strategies, and describing them provides a substrate as necessary to cognitive science as the taxonomic substrate has been to modern biology. Within cognitive science, only the linguists (and to some extent, the developmental psychologists) have had a tradition of detailed description, rather than a tradition of experimentation in search of generally valid truths.

In humans, moreover, performance programs are the product of social learning, and the programs of the twentieth century will not be identical with those of the nineteenth, or the first. Cognitive science must be prepared to study the respects in which the mind of the Greek citizen is the same or different from the mind of modern Western man, or of an African villager. There is room for a greatly increased volume of cognitive science research in this sociological direction.

The social dimension of intelligence has been recognized in social psychology, where social cognition is currently a central focus of research. Yet social psychology in general, and social cognition, in particular, has as yet had only minimal contact with cognitive science. In part, this insulation may stem from the rather special cognitive concerns that social psychology has had: especially with how people perceive others (and themselves), and how they go about forming these perceptions. Whatever the reason, one may expect that social psychology will play a larger role in cognitive science in the future than it has up to the present time.

Conclusion

Intelligence is closely related with adaptivity -- with problem-solving, learning, and evolution. A science of intelligent systems has to be a science of adaptive systems, with all that entails for the difficulty of finding genuine invariants. Some of the invariance in intelligence is imposed by the structure of the inner environment -- the limits, for example, of human short-term memory. Some of it is imposed by the outer environment -- the need to search very large spaces selectively. Some of the invariance is to be found in the structure of learning systems, rather than in the highly adapted performance systems they produce. But, in cognitive science we must be prepared to recognize that the invariants in an adaptive system are likely to be limited to specific times and places -- that in the long run, almost any aspect of them can change adaptively.

References

Abelson, R.P. (1953) Computer simulation of "hot" cognition. In S.S. Tompkins & S. Messick (Eds.) *Computer simulation of personality.* New York: Wiley.

Anderson, J.R. (1983) *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Bhaskar, R. & Simon, H.A. (1977) Problem solving in semantically rich domains: An example from engineering thermodynamics. *Cognitive Science,* 1, 193-215.

Bransford, J.D. & Franks, J.J. (1971) The abstraction of linguistic ideas. *Cognitive Psychology* 2, 331-350.

Broadbent, D.E. (1958) *Perception and communication.* New York: Pergamon Press.

Chase, W.G. & Clark, H.H. (1972) Mental operations in the comparison of sentences and pictures. In L.W. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley.

Chomsky, N. (1965) *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Coles, L.S. (1972) Syntax directed interpretation of natural language. In Simon, H.A. & Siklössy, L. *Representation and meaning.* Englewood Cliffs, NJ: Prentice-Hall.

Corsini, R.J. (1987) *Concise encyclopedia of psychology.* New York: Wiley.

D'Andrade [CHAPTER IN THIS VOLUME]

Dreyfus, H.L. (1972) *What computers can't do.* New York: Harper & Row.

Dyer, M.G. (1983) *In-depth understanding: A computer model of integrated processing of narrative comprehension.* Cambridge, MA: MIT Press.

Ericsson, K.A. & Simon, H.A. (1984) *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fisher, C. (1988) Advancing the study of programming with computer-aided protocol analysis. In G. Olson, E. Soloway, and S. Sheppard (Eds.), *Empirical studies of programmers: 1987 Workshop.* Norwood, NJ: Ablex.

Feigenbaum, E.A. & Simon, H.A. (1986) EPAM-like models of recognition and learning. *Cognitive Science* 8,305-336.

Gerloff. P. (1987) Identifying the unit of analysis in translation: some uses of think-aloud protocol data. In C. Faerch & G. Kasper (Eds.)

Introspection in second language research. Philadelphia, PA: Multilingual Matters.

Glass, G. (1983) Synthesizing empirical research: meta-analysis. In S.A. Ward and L.J. Reed (Eds.) *Knowledge structure and use: implications for synthesis and interpretation.* Philadelphia: Temple University Press.

Green, B. and Hall, J. (1984) Quantitative methods for literature review. *Annual Review of Psychology,* 35, 37-53.

Hayes, J.R. & Simon, H.A. (1974) Understanding written problem instructions. In L.W. Gregg (Ed.), *Knowledge and cognition* Potomac MD: Erlbaum.

Hebb, D.O. (1949) *The organization of behavior.* New York: Wiley.

Holmes, F.L. (1979) Hans Krebs and the discovery of the ornithine cycle. *Proceedings of the Symposium on Aspects of the History of Biochemistry,* FASEB

Just, M.A. & Carpenter P.A. (1980) A theory of reading: From eye fixations to comprehension. *Psychological Review,* 87, 329-354.

Just, M.A. & Carpenter, P.A. (1987) *The psychology of reading and language comprehension.* Boston, MA: Allyn & Bacon.

Kahneman, D., Slovic, P. & Tversky, A. (Eds.) (1982) *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kaplan, C.A. (1987) *Computer simulation: separating fact from fiction.* Unpublished manuscript. Pittsburgh: Carnegie-Mellon University.

Kaplan, C.A. & Simon, H.A. (unpublished)

Kosslyn, S.M. (1980) *Image and mind.* Cambridge, MA: Harvard UniversiPress.

Kulkarni, D. & Simon, H.A. (1988) The processes of scientific discovery: The strategy of experimentation. *Cognitive Science* (to appear).

Larkin, J.A. & Simon, H.A. (1987) Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science* 11, 65-100.

Lindzey, G. & Aronson, E. (1968) *The handbook of social psychology.* (2nd ed.) Reading, MA: Addison-Wesley.

McCarthy, J. & Hayes, P. (1969) Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.) *Machine intelligence 4.* Edinburgh: Edinburgh University Press.

Miller, G.A. (1956) The magical number seven, plus or minus two. *Psychological Review* 63, 81-97.

Miller, G.A., & Johnson-Laird, P.N. (1976) *Language and perception.*

Cambridge, MA: Harvard University Press.

Newell, A. (1980) Physical symbol systems. *Cognitive Science* 2, 135-184.

Newell, A. [SOAR CHAPTER IN THIS VOLUME]

Newell, A. & Simon, H.A. (1956) The logic theory machine. *IRE Transactions on Information Theory*. IT-2(3), 61-79.

Newell, A. & Simon, H.A. (1963) Computers in psychology. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology, I* New York: Wiley, 361 - 428.

Newell, A. & Simon, H.A. (1972) *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett R.E., & Wilson, T.D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 84, 231-259.

Norman, D.A. (Ed.) *Perspectives on cognitive science.* Norwood, NJ: Ablex.

Paige, J.M. & Simon, H.A. (1966) Cognitive processes in solving algebra word problems. in H.A. Simon (Ed.) *Models of thought.* New Haven, CT: Yale University Press.

Posner, M.I. & McLeod, P. (1982) Information processing models: In search of elementary operations. *Annual Review of Psychology* 33, 477-514.

Pylyshyn, Z. (1973) What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* 80,1-24.

Quillian, M.R. (1966) *Semantic theory.* .Unpublished doctoral dissertation, Carnegie Institute of Technology.

Rich, E. (1983) *Artificial Intelligence.* New York: McGraw-Hill.

Rochester, N., Holland, J.H., Haibt, L.H. & Duda, W.L. (1956) Test on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory.* IT-2(3),80-93.

Rosenberg, S. & Simon, H.A. (1977) Modeling semantic memory: Effects of presenting semantic information in different modalities. *Cognitive Psychology* 9,293-325.

Samuel, A.L. (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3,210-229.

Searle, J.R. (1980) The intentionality of intention and action. *Cognitive Science* 4, 47-70.

Siklössy, L. (1972) Natural language learning by computer. In

Simon, H.A. & Siklòssy, L. (Eds.) *Representation and meaning.* Englewood Cliffs, NJ: Prentice-Hall.

Simon, H.A. (1977) *Models of discovery.* Boston, MA: D. Reidel Publishing.

Simon, H.A. (1979a) *Models of Thought.* New Haven, CT: Yale University Press.

Simon, H.A. (1979b) Information processing models of cognition. *Annual Review of Psychology,* 30, 363 -396.

Simon, H.A. (1980) Cognitive science: The newest science of the artificial. *Cognitive Science* 4, 33-46.

Stone, P.J., Dunphy, D.C., Smith, M.S., & Ogilvie, D.M. (1966) *The general inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.

Thibadeau, R., Just, M.A., & Carpenter, P.A. (1982) A model of the time course and content of reading. *Cognitive Science,* 6, 157-203.

Turing, A.M. (1950) Computing machinery and intelligence. *Mind* 59, 433-460.

Waterman, D.A. & Newell, A. (1971) Protocol analysis as a task for artificial intelligence. *Artificial Intelligence,* 2, 285-318.

Waterman, D.A. & Newell, A. (1973) PAS-II: An interactive task-free version of an automatic protocol analysis system. In *Proceedings of the Third IJCAI.* Menlo Park, CA: Stanford Research Institute, 431-445.

Weber, R.P. (1985) *Basic content analysis.* Beverly HIlls, CA: Sage Publications.

Wolf, Fredric M. (1987) *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage Publications