Computational Theories of Cognition

Herbert A. Simon

Carnegie Mellon University

QUAL.MSS

Acknowledgments

This research was supported by the National Science Foundation, Grant No. DBS-9121027; and by the Defense Advanced Research Projects Agency, Department of Defense, ARPA Order 3597, monitored by the Air Force Avionics Laboratory under contract F3361 5-81-K-1539. Reproduction in whole or in part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

Computational Theories of Cognition

Herbert A. Simon
Carnegie Mellon University

In the late 1950's, the hypothesis was advanced that human thinking is information processing, alias symbol manipulation. Like most new ideas, this one had many harbingers, especially those collected under the general rubric of cybernetics. Allen Newell and I gave some account of these precursors in the historical addendum to our *Human Problem Solving* (1972). What was new, beginning about 1956, was the translation of these ideas into symbolic (non-numerical) computer programs that simulated human mental activity at the symbolic level. The traces of these programs could be compared in some detail with data that tracked the actual paths of human thought (especially verbal protocols) in a variety of intellectual tasks, and the programs' veracity as theories of human thinking could thereby be tested.

Thus the digital computer provided both a means (programs) for stating precise theories of cognition and a means (simulation, using these programs) for testing the degree of correspondence between the predictions of theory and actual human behavior. Prominent early examples of programs that were successful in matching substantial ranges of human behavior included EPAM, a program that simulates human perception and learning, and GPS, a program that simulates human problem solving. Over the years, EPAM has steadily expanded the range of tasks it handles, while GPS has been transmuted into Soar, a major step toward the unification of cognitive theories. Both are still very much alive.

Before the information-processing approach could play a substantial role in psychology, it had to overcome a number of deeply entrenched beliefs associated with Behaviorism, on the one hand, and with Gestalt psychology, on the other. Behaviorists were suspicious of attempts to theorize about what went on inside the head (although hard-core Behaviorism had already been somewhat softened on this dimension by the theories of Tolman and Hull). Gestaltists were opposed to reductionism and to mechanistic accounts of the phenomena that they regarded as "intuitive" and "insightful." There was also a third view, which insisted that "real" explanations of psychological phenomena must be physiological and neural: that there was no room for a level of symbolic theories between behavior and the biological brain. All of these views had to be reconciled with the symbolic one for the latter to gain any great credence.

Finally, in the early years, few psychologists had opportunities for hands-on interaction with computers. For most of them, the computer was a "mechanistic" number-crunching device, a black box occupied by 0-1 bits. It hardly seemed a promising candidate for representing the flexibility, fallibility and richness of human thought.

The idea was quite novel that a program was analogous to a system of differential (or difference) equations, hence could express a dynamic theory. Equally novel was the idea that computers were not confined to the numerical, but could represent symbols (patterns with denotations) of all kinds. Thinking aloud was confounded with introspection, hence not regarded as a legitimate sources of empirical data. Standard statistical techniques were not of much use for judging the goodness of fit of computer traces to verbal protocals.

The Physical Symbol System Hypothesis

Several decades passed before the foundational ideas of information processing permeated the psychological community and began to replace the established assumptions. Only in the 1970's can it be said that the new approach had gained a dominant position in cognitive psychology. Even then, there were many more psychologists who, while accepting the greater freedom in experimentation and theorizing that accompanied the revolution, were less than enraptured by the "computer metaphor," as it was often called, or the validity of verbal protocols as data. Most research in cognition continued to use conventional experimental methods and designs and most theories continued to be expressed in loose verbal terms.

It was only after many psychologists began to interact with personal computers, thereby acquiring a more sophisticated picture of what a computer is, that appreciation grew for the idea of a computer program as a theory, or for the idea of testing theories by comparing computer outputs with human verbalizations. Nor has the view of information processing psychology that I have sketched above permeated the whole profession even today. Those of us who regard computer programs as "theories" rather than "metaphors" are probably still in the minority. "Weak AI," as the metaphoric view is sometimes called, still probably has more advocates than "strong AI."

The basic assumption that underlies the strong view is the Physical Symbol System Hypothesis, which Allen Newell and I set forth in our 1975 ACM Turing Award Lecture (Newell and Simon, 1976):

A physical symbol system has the necessary and sufficient means for general intelligent action.

A *physical symbol system* (PSS) is a system that has the capacities of a modern digital computer: to input and output symbols, organize and reorganize them in symbol structures, store and erase them, compare them for identity or difference, and behave contingently on the outcome of such comparisons. *Symbols* in a PSS are simply patterns of any kind (and made of anything) that point to or denote something other than themselves. It should be emphasized that there is no assumption that the symbols in a PSS are in any sense verbal -- they can be patterns of *any* kind: verbal, pictorial, abstract, neural. In the human brain, such patterns sometimes are and sometimes are not accessible to conscious awareness. They may denote other symbols in the brain or external objects or configurations.

The PSS hypothesis makes two empirical claims:

(1) that a PSS can be programmed to behave intelligently;

(2) that human beings are intelligent by virtue of being physical symbol systems; their intelligent behavior is to be explained in terms of symbols and symbol processes.

It is easy to see the connection between these claims and the "strong" version of information processing psychology: Because the symbolic processes that account for human intelligence are all available to computers, we can write computer programs that produce intelligent behavior using processes that track closely the processes of human intelligence. These programs, which predict each successive step in behavior as a function of the current state of the memories together with the current inputs, are theories, quite analogous to the differential equation systems of the physical sciences. To test these theories, we need moment-by-moment data on human thought processes; the finest-grained data of this sort currently available are verbal thinking-aloud protocols.

During the first several decades of information processing psychology, the thought processes hypothesized were predominately serial, one-at-a-time, processes. In recent years, there has been a growing interest in parallel systems as an alternative architecture. My personal view is that there is a large place for both in the working of the brain; but a final determination of the respective roles of serial and parallel processing in human thinking lies in the future, and I will not discuss the issue in this chapter.

Accompanying the interest in parallel architectures, the popularity has increased of

systems that try to incorporate at least some of the neural organization of the human brain. These connectionist systems forego a purely symbolic level of theorizing in favor of representing cognitive activities directly by a network whose elements are somewhat nerve-like in character. The "neurons" may be quite abstract, as in most connectionist systems (Rumelhart and McClelland, 1986), or they may attempt to capture more of the physico-chemical properties of actual neurons. In either case, these systems tend to return to an older tradition that sees no need for a symbolic level of theorizing above the neurological level. Until such time as the connectionist and other network models are able to handle a range of complex cognitive tasks comparable to those already explained by the symbol-level theories, there does not seem to be much point in trying to resolve this issue.

There is also some disagreement as to whether connectionist and neural-network systems are to be regarded as symbolic (as PSS's). Having discussed this question elsewhere (Verba and Simon, forthcoming), I will not take it up here.

The body of empirical evidence in support of the PSS as a theory of human cognition (in the first of the variant forms I have described) is by now very large. One can get a general picture of it from such sources as Newell and Simon (1972), Anderson (1983), Simon (1979, 1989), Langley et al. (1987) and Newell (1990). The successful applications of the theory range all the way from classical learning experiments, to problem solving, concept attainment, learning from examples, understanding written instructions, learning natural language (Siklossy, 1972), free recall, chessplaying, visual perception and mental imagery, scientific discovery. The references cited above are only a sample, and they exclude the numerous intelligent systems in AI that are not claimed to simulate human processes.

I sometimes find it surprising (and not a little frustrating) that this empirical literature is so seldom cited, and even less frequently examined in any detail, in discussions of the validity of the PSS hypothesis. There appears still to be a widespread belief that the nature of human thought processes can be determined from first principles without examining human behavior in painstaking detail and comparing that behavior with the claims of rigorous theories. I will try to avoid that mistake in the remainder of this paper; when I make empirical claims, I will refer to the relevant empirical research.

Philosophical Import of the PSS

Before turning to more empirical matters, I should like to comment on the implications of the physical symbol system hypothesis for some classical topics in philosophy. The first of these is epistemology. The second is the mind-body problem.

Epistemology for Computers

Epistemology is concerned with the question of how, since we live, so to speak, inside our heads, we acquire knowledge of what there is outside our heads. Idealism ducks the question by locating everything of interest inside the head, thereby avoiding any problem of transport. Empiricism, in any of its forms, lacks this escape. Quine, in *Word and Object* (1960), abandoned the attempt to find out how *I* [Quine] know, and asked instead how *he* [a native informant] knows. Then everything was outside the head -- at least Quine's head -- and the difficulties of dealing with sensation and perception did not have to be faced. But they were replaced by the difficulty that the inside of the native informant's head was almost as inaccessible to Quine as the world outside his own [Quine's] head. The native informant's behavior, including his verbal behavior, was the only clue to what was going on inside his head.

As early as 1955, Rudolf Carnap (1956) had a suggestion for removing this problem. His idea was to employ an (intelligent) computer as the native informant. To determine what the computer knows, and how it came to know it, one has access not only to its behavior (outputs) but also, in any desired detail, to its inner workings and the successive states of its memory. One can use these data to construct a wholly empirical theory of how it comes to know, and what its knowledge consists in -- an epistemology for computers. Today, we can actually carry out this enterprise, and in a recent paper (Simon, 1992) I sketched out how the concept of analyticity might be explicated by this means.

But I can provide even more concrete evidence for the feasibility of this kind of exploration in epistemology. Humans progress from a state of innocence to a state in which they have some command of a natural language and the ability to use language to understand and communicate denotations relating to the world outside the user's head. The task of acquiring a language has appeared to some to be so difficult that they have postulated an innate "language capacity" to account for such acquisition. This avoids, at least in part, the necessity for solving the epistemological problem -- in this view, it was solved before birth.

However, if we could build a computer program that, starting as a neonate, actually used its

eyes, ears and brain to acquire natural language, it would inform us about the preconditions for such an achievement -- what a "language capacity" would consist in. If the program matched early language learning in humans, then it would provide an answer to that part, at least, of the epistemological question of how we know.

In the late 1960's, Laurent Siklossy built such a program (Siklossy, 1972), which consequently is at least a first approximation to an empirical theory of human language learning. The program is called ZBIE (which is not an acronym for anything). ZBIE demonstrably can learn at least the simpler parts of the vocabulary, syntax and semantics of whatever language it is exposed to, and has been tested with three or four (English, German, French, Russian).

Siklossy assumes that the child is born with, or has acquired by the age when language learning begins, the ability to build symbol structures in memory ("mental pictures") of simple situations that appear before its eyes: the family dog chasing the family cat, say. This situation would be held in memory as a symbol structure consisting of two patterns corresponding to the objects (the cat and the dog) embedded in a larger pattern corresponding to the relation (chasing) that connects them. These structures are symbols, whose denotation is the external scene. They are built up in memory by sensory and perceptual processes that have no linguistic content, but are best regarded as "pictorial." They correspond to the ability that children have, before they begin to acquire language, of recognizing common objects and situations when they see them.

Now encoded situations are presented to Siklossy's system paired with sentences in the language to be learned. The sentences are intended to denote the corresponding situations. Thus, ZBIE would be given, with the symbol structure denoting that the dog is chasing the cat, the sentence: THE DOG CHASES THE CAT. After a series of such paired stimuli have been presented, ZBIE will have stored in memory (a) a net capable of recognizing the various objects it has encountered and the various situations (relations) in which they are found, (b) a similar net capable of recognizing the English words it has encountered, (c) structures that capture the rules for organizing words of various kinds into grammatical sentences, and (d) links that match words with the objects and relations that they denote.

Suppose, for example, that ZBIE has assimilated DOG, CAT, BOY, GIRL, PETS, CHASES, SEES, and so on, and also such scenes as the dog chasing the cat, the girl chasing the boy, and the boy petting the dog, together with sentences denoting those scenes. Now a new scene, never before encountered by ZBIE, is presented: the girl petting the cat. On request, ZBIE will produce

the sentence, "THE GIRL PETS THE CAT." Thus ZBIE satisfies Chomsky's first requirement for a learner or user of natural languages: that it be capable of understanding and generating sentences that it has never previously encountered.

For a more complete account of the capabilities and limits of ZBIE see Siklossy's account (1972). Our present interest is in the light that ZBIE throws on the symbolic processes that are involved in learning a language. In particular, even this simple example serves as a refutation of Searle's "Chinese Room" argument, that computers can't understand language. In-principle "impossibility" must yield to in-fact realization. Unlike the system of Searle's Chinese Room, ZBIE possesses links between its words and the outside-world things and relations that they denote.

ZBIE also leaves some important epistemological questions unanswered. It indicates what kind of semantic information (mental images) would be needed to support language learning, but it does not indicate what sensory and perceptual processes would extract this information from the outside world. This is, of course, a question of great current research interest in artificial intelligence in general, and robotics in particular. While it has not been answered in any comprehensive sense, there exist today a number of robotic systems that build internal symbol structures representing external situations: for example, the NAVLAB system that uses its own sensory, interpretive, and motor capabilities to navigate a vehicle at speeds up to about 50 miles per hour on a highway.

So we do possess the kind of in-principle answer that we require for purposes of epistemology. We only need look at NAVLAB to see how useful and veridical mental images of external situations can be acquired and used. This does not imply that NAVLAB's way of doing this resembles the human way in any close sense; what NAVLAB demonstrates is the existence of mechanisms that can build internal representations of the outside world that are usable for guiding action.

The Mind-Body Problem

With a large number of programs in existence capable of many kinds of performances that, in humans, we call thinking, and with detailed evidence that the processes some of these programs use parallel closely the observed human processes, we have in hand a clear-cut answer to the mind-problem problem: how can matter think and how are brains related to thoughts?

The PSS hypothesis asserts that the prerequisites to thinking are patterns that can be stored and manipulated. Knowledge resides in the patterning of matter, in combination with processes that can create and operate upon such patterns. A thought about a cat, perhaps induced by looking at it or by remembering it, is a symbol structure in that part of the brain (probably the frontal lobe) where the "mind's eye" resides. The processes that operate on symbol structures at this site can extract, for example, information about substructures that might reveal the color of the cat's fur, the length of its whiskers or the presence or absence of a tail. There is nothing mysterious about all of this, for computers can and do accomplish such processes. If their present capabilities for doing so fall far short of human capabilities the discrepancy does not imply any gap in our knowledge of the fundamental principles involved.

We generally demonstrate our command of the laws of physics by performing very precise, but simple, laboratory experiments, in which we allow only a few things to vary, and lock out the rest of the world as best we can. The knowledge of the basic laws we gain by this strategy, does not guarantee the predictability, much less constructibility, of more complex phenomena. Meteorologists may be (and frequently are) unable to predict the weather, and are even more frequently unable to do anything about it. This does not reduce our confidence that the atmosphere behaves in conformity with the laws of physics. Much confused discussion about artificial intelligence could be eliminated if we applied to it the same rules for evaluating our knowledge as we use in the older sciences.

The Limits of Computation

Our discussion of epistemology and the mind-body problem has not touched on one important issue that frequently arises in debates about the validity of the physical symbol system hypothesis. Even if it were granted that certain human activities usually regarded as entailing thought can be simulated by computer, perhaps there are other, qualitatively different, activities that cannot. Cognitive simulation is sometimes claimed to work only for "toy problems" (e.g., puzzles), or "laboratory problems" divorced from the everyday world, or "well-structured" problems that do not capture the vagueness of the situations that professionals must handle. Surely our examples drawn from language learning and robotic perception go well beyond these limits, but no matter: they do not guarantee that we have covered the full territory of human thinking. There may be forms of thinking qualitatively different from those that have been

simulated.

There are at least two different lines of argument that claim to show that computer simulation cannot capture human thinking in all of its forms. The first objection is that the computer is too mechanistic and "rational." The computer cannot make mistakes ("creative" or otherwise) as people do and cannot think along the "nonlinear" paths that humans can follow, and which sometimes lead to their best ideas.

The second objection (really three related objections) is that computers are incapable of "intuitive," "insightful," or "creative" thinking. The two sets of objections are not wholly unrelated, but as they do raise somewhat different questions we will take them up in turn.

Is Computation Mechanistic and Logical?

By any reasonable definition of "mechanism," the computer is a mechanism. But most biologists and probably most cognitive psychologists would agree that the brain is also. If by a mechanism we mean a system whose behavior at a point in time is determined by its current internal state combined with the influences that simultaneously impinge upon it from outside, then any system that can be studied by the methods of science is a mechanism.

But the term "mechanism" is also used in a narrower sense to refer to systems that have the relatively fixed, routine, repetitive behavior of most of the machines we see around us. Any signs of spontaneity exhibited by our toaster, our washing machine, our automobile or the factory's machine tools are to be attributed to our actions upon them (or, in more recent times, the actions of a computer that controls them).

Clearly the computer occupies an ambiguous position here. Its behavior is more complex, by orders of magnitude, than any machine we have known; and not infrequently it surprises us, even when it is executing a program that we wrote. Yet, as the saying goes, "it only does what you program it to do." But truism though that saying appears to be, it is misleading on two counts. It is misleading, first, because it is often interpreted to mean: "It only does what you believe you programmed it to do," which is distinctly not the case.

More serious, it is misleading because it begs the question of whether computers and people are different. They are different (on this dimension) only if people behave differently from the way they are programmed to behave. But if we include in "program" the whole state of human memory, then to assert that people "don't do only what they are programmed to do" is equivalent to asserting that people's brains are not mechanisms, hence not explainable by the

methods of science. It is not computer simulation that is at stake, but the possibility of explaining behavior at all.

It is best that we put aside the treacherous terms "machine" and "mechanical," and ask more directly in what sense a programmed system can exhibit spontaneity. By "spontaneity," we denote behavior that is unpredicted, perhaps even by the behaving system. Because we have very limited capacity to predict, for more than the shortest intervals, our responses to our thoughts, much less our responses to processes that are going on subconsciously in our minds, it is not surprising that we view our behavior as having large elements of spontaneity. Nor, considering the limited knowledge others have of our mental states, is it surprising that our behavior appears even more spontaneous to them?

But we can say the same of computers. If the computer is controlled by a program of any great complexity in a problem-solving task, we are unlikely to be able to predict much about what it will do next. Many of us have had the experience of playing against a chess program and addressing it in increasingly anthropomorphic terms as it surprises and threatens us by its responses to our moves. Computers only behave "mechanistically" when they are performing dull, number-crunching tasks like inverting matrices or solving partial differential equations. Then we may be able to detect the repetitive cycles in their work. In tasks of the kinds that are problematic for people, their behavior is much less simply patterned.

We must also take into account spontaneity in both the short and the long runs. In the long run, people can learn: they can change their programs. But of course computers can also learn. The EPAM program discriminates among objects by sorting them down through a net, testing them at each step to send them down different branches. EPAM has other processes that allow it to expand the net when it discovers that it has sorted something incorrectly (two different stimuli, say, to the same node). With the expanded net, it now discriminates differently (and presumably more finely) then it did before. It has learned.

Many learning mechanisms, some of them modeled on human learning processes as EPAM's is, have been incorporated in computer programs. Among the more interesting from a psychological standpoint are *chunking* (combining pieces of information or processes into larger familiar units), and learning from examples (modifying programs by examining the steps by which problems are solved and adding new processes to match the steps detected in the examples). Chunking mechanisms provide explanations for the gradual automation of highly practiced human

skills, while learning from examples matches processes that have frequently been observed in school situations.

We conclude that computers are capable of the same kinds of spontaneity that people are capable of, and that how much spontaneity they exhibit depends on the task in which they are engaged (as is true of people) and the complexity of the program with which they address the task. To the extent that we succeed in modeling our simulation programs on the behavior of people, computers exhibit the same kinds and degrees of spontaneity as people do in the same tasks.

Similar comments can be made about the "rationality" of computers as compared with people. With care and luck, we can program a computer to add a column of figures or to multiply numbers without error. If we want to simulate human behavior, we will not do this. Instead (as has been done in the study of children's arithmetic "bugs"), we will write programs that will reproduce the kinds of errors our human subjects make, and thereby provide insight into the flaws (from a logical standpoint) of the human programs.

Let me provide a more subtle example. EPAM has recently been modified to simulate the behavior of a person who had learned (over a period of three years of daily practice!) to recall sequences of 100 numbers read to him at a rate of one digit per second. The human subject only succeeded on about half the lists; on the others he made one or two errors. To simulate the human expert, we had to construct and test a theory of the source of the human errors and match them in EPAM's performance. Furthermore, we had to incorporate the error sources in EPAM's program in such a way that EPAM would continue to match human behavior in the numerous other task environments in which it had previously been tested (Richman, Staszewski, and Simon, 1993). The goal was not to improve EPAM's performance but to understand and simulate the limits on the human performance.

Viewing the question of rationality more generally, a system is rational to the extent that its behavior is well adapted to reaching its goals without excess time or effort. Economics has attempted to explain human behavior in markets, with some degree of success, by assuming an impossibly high degree of human rationality (maximization of expected utility). The theory works, when it does, because in situations that are not too complex the behavior of an adaptive system will be closely molded by its goals and the shape of the environment it is in.

John Anderson (1990) has undertaken to show that the same process applies to some

psychological phenomena. In this case, the learning mechanism is evolution, not thought. Over long periods of time, we would expect evolutionary mechanisms to mold both the behaviors and the structures of organisms so that they will be adaptive to their environments in the light of their goals and needs.

As complexity increases, departures from perfect rationality become more and more apparent, for the organism becomes less able to compute the optimal responses. In very simple environments, we can make good predictions of the behavior of adaptive organisms simply by asking what would be the most appropriate behavior to reach the goals of adaptation. In more complex environments, we need to take account of the computational limitations of the organism -- the means, innate or learned, it has available for discovering appropriate courses of action. Cognitive psychology has identified limits in the capacity of short-term memory, along with limits on knowledge, as the two most important internal parameters that determine how, and how well, a person will adapt to a complex environment.

Human behavior is seldom perfectly rational. It is almost always boundedly rational, where the limits on its ability to find optimal paths are limits on the knowledge available to it and limits on human ability to compute the consequences of actions. In order to program computers to simulate human behavior in the face of cognitive complexity we must incorporate the same limitations in the computer simulation as we have found operative in humans. If the program is too "logical," if it is not boundedly rational, it will obviously not think as the human subjects do. That would be a failure in our programming, not an intrinsic limit on the capability of computers for simulating human thought.

To produce computer programs that are only boundedly rational is not very hard. If we propose a problem that is very large and that also lacks a simple and clean mathematical structure, it is generally not possible to produce a scheme, even for the most powerful computers, that will find an optimal solution. If we add the condition that the available information is radically incomplete and inaccurate, the boundedness of the rationality is guaranteed.

Deep Thought, the most powerful chess program today, is only boundedly rational; it does not have any way of guranteeing that it has found the moves that are best in a game-theoretical sense, although its successes in competition with human players show that it generally finds very good moves (better moves than most human players can find). If we now limit Deep Thought's computing speed and capacity to a more human level, its rationality becomes even more severely

bounded. To simulate human chess play, we try (with moderate success to date) to write programs that compensate for the computing power that is lacking in humans with selective heuristics that make the search more efficient.

The same comment applies to concerns about the "linear thinking" of computers. It is not clear whether the phrase "linear thinking" has any but a metaphoric meaning, but whether it does or not, there is nothing about computers that requires them to think either more or less "linearly" than people. All depends on the program.

Intuition, Insight, Creativity

Human thought, another argument goes, is not simply a matter of search, however selective, through a maze. There are more elegant forms of thought, and these are required in order to discover and formulate problems, to have intuitions about them and gain insights into them. These forms of thought, quite different from those that have been simulated, it is said, come into the picture when people are thinking creatively. A machine, the argument concludes, can only do mechanistic thought, and hence will fail when intuition, insight and creativity are called for.

To investigate these possibilities, we must have definitions of the key terms: terms like "intuition," "insight" and "creativity." If we can provide some criteria for judging when intuition has occurred, insight has been gained or thought has been creative, then we can inquire about the processes that produce such events.

Intuition

We usually recognize the presence of intuition when someone solves a problem quite rapidly ("instantaneously") upon presentation, and especially when he or she cannot give an account of how the solution came about. ("It just suddenly entered my mind.") You speak to your doctor, and after describing a few symptoms, she provides the Latin name of a disease (and perhaps some advice about treatment). You ask how she knows. "It's obvious; any competent doctor would recognize it immediately."

The word "recognize" is the clue. The process of solving problems intuitively is indistinguishable from the process of recognizing a familiar person or object. The recognition is sudden, and we are unable to describe exactly what features of the person or object led to it. Sometimes, we make mistakes in recognition (read: "our intuitions are fallible"). It really isn't our friend but a stranger, who doesn't even look much like the friend as he comes closer. Or, the

doctor responded with the wrong Latin word: the disease from which we are suffering has a different name, but the two diseases have some symptoms in common.

Our acts of recognition or intuition become more reliable as we become more knowledgeable, more expert. Evidence has accumulated over the past twenty years that possession of a large, richly indexed, "encyclopedia" of information about a domain is the key to expertise in that domain, and to having reliable intuitions (recognitions) about problems in the domain. (For some pointers to the extensive literature on this topic see Charness, 1989; and Ericsson and Staszewski, 1989). The index to the encyclopedia consists precisely of the cues that will be recognized when the problems present themselves to the expert. Recognition gives access to the information associated with the cue in memory.

Recognition processes are modeled in detail in the EPAM system, which provides a theory, supported by extensive empirical evidence, of expert behavior (Richman, Staszewski and Simon, 1993). Recognition processes also play a central role in most expert systems, some of which imitate at least in part the processes of the human experts they were modeled on.

We conclude that no forms of thinking beyond those already modeled in recognition systems like EPAM are required to account for human intuition, its general reliability in domains of expertise, and its frequent unreliability outside these domains.

Insight

The term "insight" is sometimes used almost synonymously with "intuition." But we also use it to refer to our depth of understanding of a situation, and especially to ways of representing the situation that yield the deeper understanding. Thus Einstein gained a new insight into the problem of relative motion when he realized (recognized?) that to synchronize clocks at different places, signals have to be sent from the one place to the other. This insight, together with his belief that the speed of light would be the same in every reference frame, led him to the Lorentz transformation and the theory of special relativity.

In a recent study, we helped students to understand special relativity by presenting the same word pictures that Einstein presented in his original 1905 paper (Qin and Simon, 1992). By drawing a simple diagram, they were able to compute the time it would take for a ray of light to go from one end of a rod to another and back, for different conditions of the rod's motion. Combining the equations for a moving and a stationary rod, they were able to set up the equation from which the Lorentz equations are derived. Einstein's word pictures (he published no diagrams in his

paper!) led them to the insight that enabled them to understand the relation between the time and space coordinates in the two reference frames.

Let us consider a less esoteric example: the mutilated checkerboard, an AI problem first proposed, I believe, by John McCarthy (Kaplan and Simon, 1990). We have a checkerboard, and 32 dominoes, each capable of covering exactly two adjacent squares of the board. Obviously, we can cover the board completely with the dominoes. Now the northwest and southeast squares of the checkerboard are removed. Can we cover the remaining 62 squares with 31 dominoes?

Laboratory subjects given this problem work on it, with growing frustration, for a very long time. They try various coverings, always with no success. Sometimes, they work on a simplified, 4x4, board (also without success). Occasionally, and almost always after several hours of failure, a subject will notice that the squares left uncovered after an unsuccessful attempt at covering the board are always the same color -- the color opposite to the color of the two squares that were removed. Very frequently, subjects who notice this fact solve the problem within two or three minutes. In their verbal protocols they say: "Oh! The mutilated board has fewer red than black squares. But each domino covers one red and one black square, so it will be impossible to cover more of the one color than of the other. The problem has no solution."

The insight here consists of seeing that only the numbers of squares of each color is important, and not the arrangement of the dominoes on the board. If there are not the same number of each color, than there is no way to cover them with dominoes, which always cover a square of each color. Again, this insight appears to be closely associated with an act of recognition: recognition that each attempted covering leaves two squares of the same color uncovered.

This interpretation can be supported by repeating the experiment with some changes: providing a board with the squares uncolored; providing a board in which alternate squares are labeled "bread" and "butter." The former manipulation removes a cue that can lead to recognizing the lack of inequality of colors on the mutilated board. The later manipulation calls attention to the parity of adjoining squares by labels that have no other meaning. As predicted, in the experiment the former manipulation reduces the probability of solving the problem in a given time; the latter increases it.

Kaplan (personal communication) has written a computer program that looks for properties that remain invariant when problem solutions are attempted. It adds such properties to the

representations of the objects in the problem (adds color to the description of the dominoes and the squares). Previously, it could only compare total number of squares to be covered with the number of dominoes. Now it compares the number of each color covered by the dominoes with the number of each color on the board, and thereby discovers the impossibility.

An important source of insight in scientific discovery is surprise. Many examples can be cited of major discoveries that began with a surprise (e.g., Roentgen, X-rays; the Curies, radium; Tswett, chromatography; Fleming, penicillin; Krebs, the urea cycle; Faraday, induction of electric current by a moving magnet; and many others). In each case, surprise led to a major insight, but what are the mechanisms of surprise and insight?

Kulkarni built a program, called KEKADA, that plans experimental strategies for attacking scientific problem{w3O!f (Kulkarni and,Omyx$;}%*zO{{x1[&Qxew3i{y{~VSimon, 1988). Starting with}{o{YW{xDu{z{U a statement (more or less precise) of the goal and the methods of experimentation available, it proposes an experiment. On the basis of the outcome of the experiment, it proposes another, and so on. It forms expectations, on the basis of previous knowledge and experience, about the outcomes of the experiments it proposes. If these expectations are violated, it experiences "surprise," and begins to plan new experiments to delineate the scope of the surprising phenomenon and the mechanisms that might produce it. Using this strategy, it succeeded in simulating Kreb's experimental program for discovering the synthesis path of urea, Faraday's for explaining the production of electric current by a moving magnet, and several others.

The mechanisms that underlie KEKADA's performance are familiar. To the extent that it possesses knowledge about a domain, KEKADA can form expectations about the outcome of experiments. Having formed expectations, it can be surprised if the expectations are disappointed. ("Accidents happen to the prepared mind." -- Pasteur) Using its expert knowledge, it can plan experiments to explicate the surprise. So the insights obtained by following up a surprise again depend on powerful capacities for recognizing familiar cues.

Programs like KEKADA teach an important methodological lesson about research on cognitive processes. Since we cannot put Krebs or Faraday in a laboratory, or even interview them at this date, how can we test whether the processes that the computer program is using resemble the processes they used to make their discoveries? There is certainly no way in which we can test such simulations with the time resolution that is available when we can take verbal

protocols.

However, In the cases of Krebs and Faraday, we do have available their laboratory notebooks, which provide a complete listing of the experiments they performed -- usually several each day. We can then compare the experiments they carried out with those proposed by KEKADA, matching them both with respect to content and temporal sequence. This gives us a substantial body of data (albeit with a time grain of days rather than minutes) for testing the theory. We can also examine the scientists' publications for their (retrospective) accounts of the reasoning they employed; and in the case of Krebs, we also have retrospective interviews that the historian of science, C. L. Holmes conducted before Krebs' death.

When data of these kinds are available, we can test our models against discoveries of great historical interest and importance. We can have some confidence that if thinking, at its most powerful and imaginative, requires intuition and insight, then those qualities must be present in the psychological events that led up to such discoveries, and present also in computer programs that simulate these events, if only on a scale of days rather than minutes.

Creativity

Creativity is perhaps an even less precise term than intuition or insight. An action or its product is regarded as creative to the extent that the product has value along some dimension (aesthetic, scientific, economic, etcetera) and to the extent that it is novel. Valuable novelty is the mark of creativity. Notice that the criterion refers to the outcome, not to the process. But perhaps there are special processes that are conducive to producing valuable novelty.

In our discussion of the theory of relativity and of the experimental strategies that KEKADA models we have already entered into the province of creativity. It would be rather eccentric to claim that Einstein, Krebs and Faraday were not creative. But let us look a little farther to see whether other aspects of creativity may have been illuminated by attempts to simulate scientific discovery.

Formulating problems and providing them with effective represntations is often mentioned as an important kind of creativity. We have already had at least a glimpse, in the example of the mutilated checkboard, of how new representations might be discovered that would make a difficult problem solvable. Closely related to new representations are new concepts. Concepts like momentum, and energy and mass were not simply "there" in nature. They emerge by dint of great human effort (in this instance, the effort of such figures as Descartes, Huygens and

Newton) in response to problems of characterizing real phenomena. What would be required to simulate the discovery of new concepts of these kinds.

We do not have to speculate about the answer to this question, because the BACON program, among others, already has the capability of constructing new theoretical concepts (concepts denoting things that are not directly observable) in the course of building theories to describe data (Langley, et al., 1987). BACON is a data-driven law discovery system. Given some data from experiment or observation, it seeks to find an algebraic law that will describe the data parsimoniously. Given data on the masses and temperatures of two vials of water, and the equilibrium temperature when they are mixed together, it arrives at Black's law, which says that the equilibrium temperature is the average of the temperatures of the two component liquids, weighted by their respective masses.

But what if the liquids mixed are different, water and alcohol, say? With a little more trouble, BACON will discover that the equilibrium temperature is still a weighted average of the temperatures of the components, but the weights are now not simply the masses, but the masses multiplied by a characteristic constant for each liquid (say, *w* for water and *a* for alcohol). These constants, first discovered by Joseph Black, and subsequently and independently by BACON, are known as the *specific heats* of the respective substances.

From this example, we see that the enrichment of a problem representation by the introduction of new concepts can also be simulated. We do not have any very good evidence, as yet, that the processes that BACON uses to accomplish this are close to the processes used by human discoverers, but BACON'S processes are built on to rather simple heuristics quite similar to the heuristics we have observed humans using in other situations.

Is Natural Language Different?

Almost all of the examples I have provided of the simulation of intuition, insight and creativity have been drawn from scientific domains. People who are not scientists are sometimes reluctant to believe that these processes occur in such domains, for science is supposed to be orderly and "logical." Those of us who spend our lives in science know better, but perhaps it would be useful to allay doubts by turning to a quite different domain in which people can exhibit their xDintuition, insight and creativity. Let me ask how we might go about simulating human thinking in reading a text.

Newton) in response to problems of characterizing real phenomena. What would be required to simulate the discovery of new concepts of these kinds.

We do not have to speculate about the answer to this question, because the BACON program, among others, already has the capability of constructing new theoretical concepts (concepts denoting things that are not directly observable) in the course of building theories to describe data (Langley, et al., 1987). BACON is a data-driven law discovery system. Given some data from experiment or observation, it seeks to find an algebraic law that will describe the data parsimoniously. Given data on the masses and temperatures of two vials of water, and the equilibrium temperature when they are mixed together, it arrives at Black's law, which says that the equilibrium temperature is the average of the temperatures of the two component liquids, weighted by their respective masses.

But what if the liquids mixed are different, water and alcohol, say? With a little more trouble, BACON will discover that the equilibrium temperature is still a weighted average of the temperatures of the components, but the weights are now not simply the masses, but the masses multiplied by a characteristic constant for each liquid (say, *w* for water and *a* for alcohol). These constants, first discovered by Joseph Black, and subsequently and independently by BACON, are known as the *specific heats* of the respective substances.

From this example, we see that the enrichment of a problem representation by the introduction of new concepts can also be simulated. We do not have any very good evidence, as yet, that the processes that BACON uses to accomplish this are close to the processes used by human discoverers, but BACON'S processes are built on to rather simple heuristics quite similar to the heuristics we have observed humans using in other situations.

Is Natural Language Different?

Almost all of the examples I have provided of the simulation of intuition, insight and creativity have been drawn from scientific domains. People who are not scientists are sometimes reluctant to believe that these processes occur in such domains, for science is supposed to be orderly and "logical." Those of us who spend our lives in science know better, but perhaps it would be useful to allay doubts by turning to a quite different domain in which people can exhibit their xDintuition, insight and creativity. Let me ask how we might go about simulating human thinking in reading a text.

We see here a whole series of recognitions evoking associations already stored in memory. Although one idea suggests another, there is very little here that a logician would recognize as "reasoning." That if gin is mentioned, we are in a bar? It would be more accurate to say that bar is *suggested* by gin than that bar is *inferred* from it.

What of the author? What can we say of his mental processes? Without more data than this short passage, only some conjectures. Presumably, he began with a goal, perhaps of providing a setting and a mood for the tale he is about to tell. His memory provides him with a concrete description of a particular kind of establishment (very likely he draws from his own experience). He selects from his information store a series of words and phrases that will evoke from the reader the image of a bar, possibly in a lower-class neighborhood. He is counting on these words evoking somewhat the same associations in the reader's mind as are present in his own mind. Either he has some sense of what his readers know or he assumes that their knowledge stores resemble his.

I will not try to carry the analysis further. But perhaps it already suggests that the processes we are observing here, of both reader and writer, are not unlike the processes we have seen in other kinds of thinking, particularly those processes of recognition that we have associated with "intuition" and "insight." Just as we may "trick" a subject into solving the mutilated checkerboard problem by presenting a cue that draws attention to the alternating colors of squares on the board, so Camus "tricks" a reader into assigning personal traits to a speaker by his manner of describing him or reproducing his speech.

A brief, sketchily analysed example is not a demonstration. But I hope my discussion of the passage from *La Chute* at least raises the possibility that mental activity, quite creative mental activity, outside the sciences may yield to the same analysis as the more thoroughly studied scientific domains that I have drawn upon for my other examples.

Conclusion

In this chapter I have explored some of the central characteristics of a computational theory of cognition, and particularly the kind of theory that has been associated with the physical symbol system hypothesis. I have sought to illustrate how a computational model can be used to address epistemological questions in philosophy and the mind-body problem.

My principal focus, however, has been on the claim that human thinking can be

epistemology, *in* Burkholder, L. (Ed.), Boulder, CO: Westview Press.

Verba, A. and Simon, H. A. (Forthcoming, *Cognitive Science*) . *Reply to Touretsky and Pomerleau: Reconstructing physical symbol systems*