Explaining the Ineffable:

AI on the Topics of Intuition, Insight and Inspiration[1]

Herbert A. Simon

Departments of Computer Science and Psychology
Carnegie Mellon University

(9 September 1998)

Abstract

Artificial intelligence methods may be used to model human intelligence or to build intelligent (expert) computer systems. AI has already reached the stage of human simulation where it can model such "ineffable" phenomena as intuition, insight and inspiration. This paper reviews the empirical evidence for these capabilities, and some of their implications for the mind-body problem and related questions in philosophy.

Computer Programs as Theories

During the 1930s and '40s, and into the early '50s, I carried my Diogenes' lantern through many fields of mathematics seeking the right tools for studying human thought, but neither analysis nor finite math seemed to fill the bill. To use these mathematical tools, one had to force the phenomena into the Procrustean bed of real numbers or algebraic and topological abstractions that seemed to leave much of the content behind. Computer languages, with their ability to handle symbols of every kind, changed all that by permitting one to implement a very literal

[1] This is a modified version of a paper of the same title, published in the *14th IJCAI Proceedings* , Vol. 1, 1995, 939-948, expanded to include a discussion of the mind-body problem.

representation of human symbol processing in the machine's memories and processes.

Computer programs written in whatever languages are, at the most abstract level, simply systems of difference equations, with all of the power of such equations to describe the states and temporal paths of complex symbol systems. To be sure, these equation systems can almost never be solved in closed form; but the computer itself, in providing the powerful tool of simulation, offers a solution to that problem too.[2]

As you are well aware, the requirements of simulating the behavior of physical symbol systems called for symbol-manipulating languages quite different from the algebraic languages used in numerical computing, and led to the invention of list processing languages like the IPL's and then LISP, and still later to production-system languages like OPS-5 and logic-programming languages like PROLOG. With these languages the computer simulation can produce symbolic outputs that can be compared directly, and with very little translation, with human outputs, especially verbal protocols.

Artificial Intelligence and Cognitive Psychology

My interest in AI has been, from the beginning, primarily an interest in its application to psychology. Equally exciting opportunities emerged at the same time for designing computer programs that, without necessarily imitating human methods, could perform difficult tasks at expert professional levels. As the construction of expert systems has played second fiddle to human simulation in my own research program, I shall have little to say about it here. My focus will not be on computer *achievement* of humanoid skills, but on computer *imitation* of the processes people use to manifest such skills.

In this research, the computer program is not a "metaphor" but a precise language of theory for cognitive psychology in the same sense that differential equations are a language of theory for physics. Theories written in AI list processing languages are tested in exactly the same way as theories written in

[2] Simulation is increasingly employed within traditional mathematics as well, for the increasingly complex systems under study there also defy closed solution.

differential equations. We use the theories to make predictions, which are then tested against behavior captured in the laboratory or observed in the field.[3]

Psychology is an empirical science. It is the study of how human beings behave and of the processes occurring in their minds (that is, their brains) that bring this behavior about. The science of psychology proceeds by observing the phenomena of thinking, by building theories to describe and explain the phenomena, and by laying phenomena and theory side by side to see how closely they match. The preceding three sentences would be no more and no less true if for "psychology" we substituted "physics" or "geology" or "biology," with corresponding changes in the names of the phenomena studied. We will later describe the comparison process in more detail.

The fact that psychology is studied by scientists who themselves are human beings is of no more account than the fact that physics is studied by scientists who consist of atoms or that biology is studied by scientists who eat, breathe and procreate. What we are interested in, in all of these cases, are not the scientists but the phenomena and the theories that describe and explain the phenomena. At the general level, good methodology in physics or chemistry is good methodology in psychology. At more specific levels, each field has to invent methods and instruments for observing and theorizing that are appropriate to the phenomena of interest. The methods are to be judged by the same standards in every case.

I feel obliged to repeat these rather obvious sentiments here because books, written in armchair comfort, continue to be published from time to time that try to evaluate by philosophical means psychological theories written in computer languages. L et me explain why I regard such books as misguided. In fact, instead of trying to use philosophical analysis to settle psychological questions, which are empirical matters, I propose to reverse directions and to suggest that, with recent advances in psychology, we are now in a position to use psychological theories, and

[3] The theories of physics consist not only of the differential equations, but also certain properties of these equations that can be deduced from them (e.g., the principle of conservation of energy in mechanics). Theories defined by difference equations (programs) may also possess deducible properties, which then become part of the theory. For example, from the short-term memory structure embodied in recent versions of EPAM, the short-term memory capacity can be deduced from the structure and parameters of the program.

the empirical evidence on which they are founded, to settle some issues that have been important, historically, in philosophy.

Cognitive Psychology's Empirical Base

As psychology is an empirical science, we can only judge whether and to what extent particular theoretical proposals are valid by comparing them with data. In the face of such comparisons, philosophical speculation is superfluous; in the absence of such comparisons, it is helpless. Therefore, if we wish to evaluate the claims of theories of thinking (whether these theories take the form of computer programs or some other form), we would do well to spend most of our time studying the empirical evidence and making the explicit comparisons with the computer traces.

By now, such evidence is voluminous. This is not the place to review it, but I'll cite just one very specialized example. In the book, *Protocol Analysis* (1993), that Anders Ericsson and I have written, treating the methodology for testing cognitive theories by comparing human think-aloud protocols with computer traces, there are 42 pages of references. It is not unreasonable to ask anyone who proposes to evaluate the validity of verbal reports as data either to become acquainted with a substantial portion of this literature or to announce clearly his or her amateur status. Similarly, it is not unreasonable to ask anyone proposing to pronounce on memory capacity or the acquisition and response speeds of human memory to become acquainted with that large literature.

There are, of course, comparably large literatures on problem solving, reasoning, perceiving, and many other topics. Any serious assessment of our knowledge of human thought processes or of the veridicality of theories that purport to describe or explain these processes must rest on the data reported in this literature.

What theories are available for testing, and what kinds of phenomena do they address? Again, I can only cite a few examples, some from my own work and some from the work of others. An early example is the General Problem Solver (GPS), whose central mechanism, means-ends analysis, has been shown empirically, in numerous studies, to be a much-used heuristic in human problem solving. (A small fraction of these empirical tests are discussed in Newell & Simon, 1972; you will find others in the two volumes of my *Models of Thought*, 1979, 1989.). Contemporary with

GPS is EPAM, a model of human perceptual and memory processes due originally to Feigenbaum, which has been tested successfully against empirical data from experiments on verbal learning, expert memory performances in several domains of expertise (including expertise in mnemonics), and concept attainment. (For some of the empirical tests see Feigenbaum & Simon, 1984; and Richman, Staszewski and Simon, 1995.)

A somewhat later system is John Anderson's ACT* (1983), which focuses especially on semantic memory and the explanation of contextual effects through spreading activation. A very different and still newer theory, or set of theories, are "neural" networks of the connectionist variety that have shown capacities to learn in a variety of tasks (McClelland & Rumelhart, 1986). Quite recently, Allen Newell, in collaboration with John Laird, Paul Rosenbloom and others, has produced Soar, a vigorous push from GPS into a far more general and unified architecture, which demonstrates the relevance of multiple problem spaces and learning by chunking (Newell, 1990). Still closer to the topics I shall address in the remainder of this talk is the BACON system (see Langley, et al., 1987) and its close relatives, GLAUBER, STAHL, KEKADA (Kulkarni & Simon, 1988), LIVE (Shen, 1994) and others that simulate many of the discovery processes that are discernible in the activities of scientists. Some of the models I have mentioned are complementary, some are competitive, as theories are in any science.

To understand these systems, not just as interesting examples of artificial intelligence but as theories of human thinking, and to adjudicate among them when they conflict, we must devote just as much attention to the experimental and other empirical evidence about the phenomena they model as to the structures and behaviors of the programs themselves. Errors in the evaluation of these programs as psychological theories are caused less often by lack of knowledge or inaccurate knowledge about the programs than by lack of knowledge or inaccurate knowledge about how human subjects behave when they are confronted with the same tasks as the programs were tested on.

For one example, the brittleness of computer programs when they wander outside the task domain for which they are programmed is often mentioned as a defect of these programs, viewed as psychological theories, without noticing the extraordinary brittleness of human behavior when it wanders outside the arena of the actor's experiences. (Inexperienced urbanites lost in a wilderness frequently freeze or starve to death in circumstances where experienced savages survive. Novices playing their first bridge hand bid and discard almost randomly.) Theories cannot be compared with facts unless the theories are specified precisely and the facts known thoroughly.

Limits of Explanation?

In the remainder of this paper I shall put the information processing explanation of thinking to what is usually regarded as a severe test. The idea that the processes humans use in everyday, relatively routine and well-structured tasks can be modeled accurately by computers has gained, over the years, a considerable amount of acceptance — more among experimental psychologists than among people who are more distant from the data. The idea that these models can be extended to ill-structured tasks of the kinds that require ingenuity, perhaps even creativity, when performed by humans is less widely accepted. This is no more a philosophical question than the questions that I have discussed previously. It is a question about certain kinds of human behavior and whether these kinds of behavior can be modeled by computers. It is to be settled by comparing the records of human behavior with the output of computer models, just as we settle questions in physics by comparing the laboratory behavior of physical systems with the differential equations of physical theory.

I shall focus on three terms that appear frequently in the literature and in popularized psychology (not always with the same meanings) and which have been used to label behaviors that are often claimed to be beyond explanation by programmable mechanisms. The three terms are "intuition," "insight" and "inspiration." In addressing the cognitive phenomena associated with each of these terms, I shall first define the term so that we can determine when the corresponding phenomena are being exhibited. Without clear tests that enable us to identify the occasions of "intuition," "insight" and "inspiration," there are no phenomena to explain.

I cannot claim that the definitions I shall propose represent the only ways in which these terms are, or can be, used. I will claim that they correspond to the usual meanings, and that the operational tests on which they are based are the operational tests that are commonly used to determine when people are being "intuitive," "insightful," or "inspired." These are the properties the definitions should possess if they are to be used in theories of intuition, insight and inspiration.

Having established operational tests for the phenomena, we shall look at the evidence as to whether people and computers exhibit the processes in question, and if so, under what circumstances. What I shall show is, first, that the presence or absence of phenomena like these, sometimes claimed to be ineffable, can be determined objectively, and second, that certain computer programs are mechanisms that exhibit these phenomena and thereby provide explanations for them.

Intuition

Let me start with the process of human thinking that is usually called "intuition." Before we can do research on intuition, we have to know what it is; in particular, we must have some operational definition that tells us when intuition is being exhibited by a human being and when it is not. It is not too difficult to construct such a definition.

The marks that are usually used to attribute an intelligent act (say, a problem solution) to intuition are that: (1) the solution was reached rather rapidly after the problem was posed, and (2) the problem solver could not give a veridical account of the steps that were taken in order to reach it. Typically, the problem solver will assert that the solution came "suddenly" or "instantly." In the few instances where these events have been timed, "suddenly" and "instantly" turn out to mean "in a second or two," or even "in a minute or two."

That's essentially the way my dictionary defines intuition, too: "the power or facility of knowing things without conscious reasoning." Let us take the criteria of rapid solution and inability to report a sequence of steps leading up to the solution as the indications that people are using intuition. These are the criteria we actually use to judge when intuition is being exhibited. Applying these criteria, we now have some clearly designated phenomena to be explained; we can try to construct some difference equations (computer programs) that behave intuitively.

Intuitive thinking is frequently contrasted with "logical" thinking. Logical thinking is recognized by being planful and proceeding by steps, each of which (even if it fails to reach its goal) has its reasons. Intuitive thinking, as we have seen, proceeds by a jump to its conclusions, with no conscious deliberateness in the process. But intuitive and logical thinking can be intermingled. The expert, faced with a difficult problem, may have to search planfully and deliberately, but is aided, at each stage of the search, by intermediate leaps of intuition of which the novice is incapable. Using what appear to be (in systems programming terms) "macros," frequent intuitive jumps, the expert takes long strides in search, the novice numerous tiny steps.

A Theory (Computer Model) of Intuition

Having specified how we will recognize intuition when it occurs, the next task in building a theory of it is to design a computer program (or find one already built) that will solve some problems intuitively — as determined by exactly the same criteria as we employ to determine when people are using intuition. The program will solve these problems, if they are easy, in a (simulated) second or two and will be unable to provide a (simulated) verbal report of the solution process. Fortunately, at least one such program already exists: the EPAM program (Richman, Staszewski & Simon, 1995), which first became operative about 1960. It was not designed with intuition in mind, but rather to simulate human rote verbal learning, for which there already existed at that time a large body of empirical data from experiments run over the previous 70 years. EPAM accounted for the main phenomena found in these data.

The core of EPAM is a tree-like discrimination net that grows in response to the stimuli presented to it and among which it learns to discriminate, and a short-term memory that will hold a few familiar symbols (7±2?), but will retain them more than 2 seconds only if it has time to rehearse them. EPAM's discrimination net is somewhat similar to the Rete nets that are used to index production systems. EPAM learns the correct discriminations by experience, with only feedback of "right" or "wrong" to its responses. EPAM nets have been taught to discriminate among more than $3x10^5$ different stimuli, and there is nothing final about that number.

These learned patterns, once acquired, can now be recognized when presented to EPAM because it sorts them through its net, the recognition time being logarithmic in the total number of stimuli in the net. If the net has a branching factor of 4, then recognition of a net discriminating among a million stimuli could be achieved by performing about ten tests (4^{10} = 1,048,576). The EPAM model, its parameters calibrated from data in verbal learning experiments, can accomplish such a recognition in a tenth to a fifth of a second. If we add additional time for utterance of a response, the act of recognition takes a second or less.

Now suppose we confront EPAM with a situation that is recognizable from its previous experience (a collection of medical symptoms, say). It can now access, in less than a second, information about a disease that is presumably responsible for these symptoms. As EPAM is able to report symbols that reach its short-term memory (where the result of an act of recognition is stored), it can report the name of the disease. As it cannot report the results of the individual tests performed on the symptoms along the path, it cannot describe how it reached its conclusions. Even if it can report the symptoms that were given it (because it stored some of them in memory during the presentation), it cannot give a veridical account of which of these were actually used to make the diagnosis or how they were considered and weighed during the recognition process.[4] We might add, "even as you and I," for these are also the characteristics of human diagnosis: the physician can report what disease he or she has recognized, but cannot give a veridical report of which symptoms were taken into account, or what weights were assigned to them.

To simulate the diagnostic process in more complex cases, we need a system that contains, in addition to EPAM's discrimination net and the long-term memory it indexes and accesses, some capabilities for solving problems by heuristic search — a combination of EPAM with a sort of General Problem Solver (GPS) or Soar. Then we will observe this combined system not only recognizing familiar symptoms and their causes, but also reasoning to infer what additional tests might discriminate among alternative diagnoses that have been recognized as possible causes of the initial symptoms.

Automatic medical diagnosis systems now exist that perform diagnostic tasks far more accurately than EPAM alone could, for they take into account alternative diagnoses, do some simple reasoning about relations among symptoms, and are able to request additional tests on the patient to achieve greater discriminatory power and accuracy. These systems, of course, are using a combination of intuition, as usually defined, and "logical" thought (including means-ends analysis in some form). Our

[4] This does not mean that EPAM cannot be programmed to trace its steps, but that the simulation of its verbal processes will report only symbols that are stored, at the time of reporting, in short-term memory. The trace of non-reportable processes must be distinguished from the simulation of processes the theory claims to be reportable.

current interest is not in machine competence in medical diagnosis but in models of intuition. EPAM, as described, is exhibiting intuition, as defined operationally, and modeling at least the first stage of thought (the recognition stage) of an experienced physician confronted with a set of symptoms.

Testing the Model of Intuition as Recognition

What grounds do we have for regarding this basic recognition mechanism, which lies at the core of EPAM, as a valid theory of the process that causes people to have intuitions? Simply that it has the same manifestations as human intuition: it occurs on the same time scale accompanied with the same inability to explain the process. Nor was it explicitly "cooked up" to exhibit these properties: they are basic to a system that was designed with quite other simulation tasks in mind. This is exactly the test we apply in validating any theory: we look at the match between the theory and the phenomena and at the ratio of amount of data explained to number of parameters available for fitting.

We can extend the tests of this theory of intuition further. It is well known that human intuitions that turn out to be valid problem solutions rarely occur to humans who are not well informed about the problem domain. For example, an expert solving a simple problem in physics takes a few computational steps without any pre-planning and reports the answer. The recorded verbal protocol shows the steps, but no evidence of why they were taken (no mention of the goals, operators, the algebraic expressions in which numbers were substituted). A novice solving the same problem works backwards from the variable to be evaluated, explicitly stating goals, the equations used and the substitutions in the equations. In one experiment, the novice's protocol was approximately four times as long as the expert's (Simon & Simon, 1978) and exhibited no intuition — only patient search. Novices who replace this search by guessing seldom guess correct answers. This is exactly what EPAM predicts: that there is no recognition without previous knowledge, and no intuition without recognition. Notice that intuition can be as fallible as the recognition cues on which it is based.

There are a number of experimental paradigms for carrying out tests on this theory that intuition is simply a form of recognition. The expert/novice paradigm has already been mentioned: experts should frequently report correct intuitive solutions of problems in their domain, while novices should seldom report intuitions, and if they report any, a large proportion should be incorrect. Experts who are able to report intuitions in their domains should be unable to do so in domains where they are not expert. By making cues more or less obvious, it should be possible to increase or decrease the frequency of correct intuitions; misleading cues should induce false intuitions. Hints of various kinds should draw attention to cues, hence facilitate intuition. These are only the most obvious possibilities, all of which have been tested with positive outcomes for the theory.

Experiments on intuition are best carried out on tasks where the correctness of answers can be verified, at least after the fact. We would want to identify "false intuition" to explain the cases (probably very frequent but hard to pinpoint in domains where objective criteria of correctness are lacking) where the presence of certain features in a situation leads subjects to announce a sudden solution although the connection between the cue and the inferences drawn from it is invalid. Determining the circumstances that encourage or discourage false intuition would involve research on the characteristics of situations that subjects attend to, and the beliefs they hold that lead them to the erroneous solutions. Some of the research that has been done on the psychology of so-called "naive physics" fits this general paradigm, as does some of the research on "garden paths" (spontaneous but erroneous interpretations) in syntactic analysis of sentences.

We see that intuition, far from being a mysterious and inexplicable phenomenon, is a well known process: the process of recognizing something on the basis of previous experience with it, and as a result of that recognition, securing access in long-term memory to the things we know about it. What subjects can report about the origins of their intuitions, and what they can't report, are exactly what we would predict from a theory that explained the phenomena associated with recognition. As a matter of fact, we could simplify our vocabulary in psychology if we just abandoned the word "intuition," and used the term "recognition" instead.

Insight

Another process of thought that has sometimes been declared to be inexplicable by mechanical means is insight. My dictionary, this time, associates insight closely with intuition. In fact, its second definition of "intuition" is: "quick and ready insight." Its explicit definition of "insight" is not much more helpful: "the power or act of seeing into a situation: understanding, penetration." Again, we gain an impression of suddenness, but in this case accompanied by depth. Perhaps we shall want to regard any instance of insight as also an instance of intuition, in which case our work is already done, for we have just proposed a theory of intuition. Let's see, however if there is an alternative — some other phenomenon that needs explanation and to which we can attach the word "insight."

Consider the "aha" phenomenon. Someone is trying to solve a problem, without success. At some point, a new idea comes suddenly to mind — a new way of viewing the problem. With this new idea comes a conviction that the problem is solved, or will be solved almost immediately. Moreover, the conviction is accompanied by an understanding of why the solution works. At this point we hear the "aha," soon followed by the solution — or occasionally by a disappointed realization that the insight was illusory. In some cases, after a problem has been worked on for some time without progress, it is put out of mind for a while, and the "aha" comes unexpectedly, at a moment when the mind was presumably attending to something else.

In both scenarios, with and without the interruption, the phenomenon shares the characteristics of intuitive solution: suddenness of solution (or at least of the realization that the solution is on its way), and inability to account for its appearance. The process differs from intuition in that: (1) the insight is preceded by a period of unsuccessful work, often accompanied by frustration, (2) what appears suddenly is not necessarily the solution, but the conviction of its imminence, (3) the insight involves a new way of looking at the problem (the appearance of a new problem representation accompanied by a feeling of seeing how the problem works) and (4) sometimes (not always), the insight is preceded by a period of "incubation," during which the problem is not attended to consciously, and occurs at a moment when the mind has been otherwise occupied. The third of these features is the source of the feeling of "understanding" and "depth" that accompanies the experience of insight. Again, these are the phenomena we use to identify instances of insight in human beings (ourselves or others). We can take the presence of these four features as our operational definition of insight, and using it, we now have some definite phenomena that we can study and seek to explain.

A Theory (Computer Program) of Insight

Let me now describe a computer program that can experience insight, defined in the manner just indicated. I shall present this theory a little more tentatively than the theory of intuition proposed earlier because, while it demonstrates that a computer program can have insights, the evidence is a little less solid than for intuition that it matches all aspects of the human experience of insight.

Again, a program that combines the capabilities of EPAM and the General Problem Solver constitutes the core of the theory. (1) We suppose that a GPS-like or Soar-like problem solver is conducting, unsuccessfully so far, a heuristic (selective) search for a problem solution. (2) It holds in long-term memory some body of information about the problem and knowledge of methods for attacking it. (3) Unfortunately, it is following a path that will not lead to a solution (although of course it is unaware of this). (4) We assume that the search is serial, its direction controlled by attentional mechanisms that are represented by the flow of control in the program. (5) Much of this control information, especially information about the local situation, is held in short-term memory, and is continually changing. (6) At the same time, some of the more permanent features of the problem situation are being noticed, learned, and stored in long-term memory, so that the information available for problem solution is changing, and usually improving. (7) The control structure includes an interrupt mechanism which will pause in search after some period without success or evidence of progress, and shift activity to another problem space where the search is not for the problem solution but for a different problem representation and/or a different search control structure. (8) When search is interrupted, the control information held in short-term memory will be lost, so that if search is later resumed, the direction of attention will be governed by the new representation and control structure, hence may lead the search in new directions. (9) As the non-local information that has been acquired in long-term memory through the previous search will participate in determine the search direction, the new direction is likely to be more productive than the previous one.

Empirical Tests of the Theory of Insight

Now we have introduced nine assumptions to explain the insight that may occur when the search is resumed, which hardly looks like a parsimonious theory. But these assumptions were not introduced into the composite EPAM-GPS to solve this particular problem. All are integral properties of these systems, whose presence is revealed by many different kinds of evidence obtained in other tasks.

One body of evidence supporting this model of insight comes from an experimental investigation of the Mutilated Checkerboard problem that Craig Kaplan and I conducted a few years ago (Kaplan & Simon, 1990). We begin with a chessboard (64 squares) and 32 dominos, each of which can cover exactly two squares. Obviously, we can cover the chessboard with the dominos, with neither squares nor dominos left over. Now, we mutilate the chessboard by removing the upper-left and lower-right corner squares, leaving a board of 62 squares. We ask subjects to cover it with 31 dominos or to prove it can't be done.

This is a difficult problem. Most people fail to solve it even after several hours' effort. Their usual approach is to attempt various coverings as systematically as possible. As there are tens of thousands of ways to try to cover the board, after some number of failures they become frustrated, their efforts flag and they begin to wonder whether a covering exists. Increasingly they feel a need to look at the problem in a new way, but people seem not to have systematic methods for generating new problem representations. Some subjects simplify by replacing the 8¥8 board with a 4¥4 board, but this does not help.

Hints do help. Although few subjects solve the problem without a hint, many do with a hint, usually in a few minutes after the hint is provided. For example, the experimenter may call attention to the fact that the two squares left uncovered after an unsuccessful attempt are always the same color, opposite to the color of the excised corner squares. Attending to this fact, subjects begin to consider the number of squares of each color as relevant, and soon note that each domino covers a square of each color. This leads quickly to the inference that a set of dominos must always cover the same number of squares of each color, but that the mutilated board has more squares of the one color than of the other: Therefore, a covering is impossible.

Subjects who discover this solution, with or without a hint, exhibit behaviors that satisfy our definition of insight. The solution is preceded by unsuccessful work and frustration; it appears suddenly; it involves a new representation of the problem that makes the problem structure evident. The subjects come to the solution quite quickly once they attend to the critical property (equality of the numbers of squares of each color that are covered). This is also true of the few subjects who solve the problem without being given a hint. These subjects have their "aha!" when they attend to the fact that the uncovered squares are always the same color, and that the mutilated board has more squares of that color than of the other. Aided by cues or not, successful subjects often (literally) say "aha!" at the moment of recognizing the relevance of the parity of squares of the two colors.

Moreover, the mechanisms that bring about the solution are those postulated in our computer theory of insight, as can be seen by examining the list given above. Steps 6 through 9 are the critical ones. In the case of hints, attention is directed to the crucial information by the hint, this information is stored in memory, and the search resumes from a new point and with a new direction of attention that makes the previous attempts to cover the board irrelevant. In the case of subjects who solve without a hint, the direction of attention to the invariant color of the uncovered squares may derive from a heuristic to attend to *invariant* properties of a situation — the properties that do not change, no matter what paths are searched in solution attempts.

There are probably several such heuristics (surprise is another one) that shift peoples' attention to particular aspects of a problem situation, thereby enabling the learning of key structural features and redirecting search. The evidence for such heuristics is not limited to laboratory situations; the role of the surprise heuristic in scientific discovery has been frequently noted. I shall return to it later.

The role of attention in insight receives further verification from a variant on the experiment. Different groups of subjects are provided with different chessboards: (1) a standard board, (2) a ruled 8¥8 matrix without colors, and (3) an uncolored matrix with the words "bread" and "butter" ("pepper" and "salt" will do as well) printed on alternate squares. More subjects find the solution in condition 3 than in condition 1; and more in condition 1 than in condition 2. The reason for the latter difference is obvious: presence of the alternating colors provides a cue to which a subject's attention may be directed. What is the reason for the superiority of "bread" and "butter" over red and black? Subjects are familiar with standard chessboards and have no reason to think that the color has any relevance for this problem, hence don't attend to it. In the case of "bread" and "butter," the subjects' attention is attracted to this unusual feature of the situation; they wonder why "those crazy psychologists put those labels on the squares." Here we obtain direct support for the hypothesis that direction of attention to the key features of the situation provides the basis for solution. Noticeability of a feature is essential, whether it is provided by an explicit clue or some other means.

Incubation

The checkerboard experiments do not say anything about incubation, or whether interruption of the solution process for a shorter or longer period may contribute to solution. Here I can point to another set of experiments carried out by Kaplan (1989). He defines incubation as "any positive effect of an interruption on problem solving performance," and lists seven explanations that have been offered for it: "unconscious work, conscious work that is later forgotten, recovery from fatigue, forgetting, priming, maturation and statistical regression (p. 1)." Kaplan then carries out experiments to show, or to confirm, that (1) interruption of certain kinds of tasks (so-called divergent-thinking tasks) improves subsequent performance (i.e., incubation can be demonstrated experimentally), (2) answers supplied after an interruption differ more from the just-previous answers than do successive answers supplied without interruption (i.e., incubation can break "set"), (3) interruptions combined with a hint increase the effects of incubation (the hint shifts attention from continuing search to changing the representation), (4) hints may work without subjects' conscious awareness of their connection with the unsolved problem, and (5) subjects underestimate the time they spend thinking about the problem during an interruption. Details can be found in the original study.

Kaplan then proposes a model, which he calls a Generic Memory Model, to account for these phenomena. The model is compatible with the one we have already proposed, with the addition of so-called priming mechanisms of the kind that Quillian (1966) and Anderson (1983) incorporate in their models of semantic memory.[5] The priming mechanisms increase the probability that subjects will attend to items that have been cued, at the same time rapidly decreasing attention to items in STM and slowly decreasing attention to items in LTM. The model accounts for the fact, as the

[5] In order to explain some quite different phenomena, priming mechanisms have also been added to the most recent version of the EPAM theory.

previous model does not, that the length of the interruption is important. Neither model needs to postulate unconscious work on the problem during interruption to account for incubation.[6] Forgetting in short-term memory of information that holds attention to an unproductive line of search, and redirection of attention from search in the original problem space to search for a new problem representation are the key mechanisms in both models that account for the bulk of the empirical findings.

On the basis of the evidence I have described and the models that have been offered to explain this evidence, I think it fair to claim that there exists a wholly reasonable and empirically supported, theory of incubation, as it is observed in human discovery, that calls only on mechanisms that are already widely postulated as components of standard theories of cognition. The process of incubating ideas is as readily understandable as the process of incubating eggs.

Inspiration (alias Creativity)

The term "inspiration" is surrounded by an aura of the miraculous. Interpreted literally, it refers to an idea that is not generated by the problem solver, but is breathed in from some external, perhaps heavenly, source. To inspire, says my faithful dictionary, is to "influence, move, or guide by divine or supernatural inspiration." A bit circular, but quite explicit about the exogenous, non-material source. A Greek phrase for it was more vivid: to be inspired (e.g., at Delphi) was to be "seized by the god."

The notion that creativity requires inspiration derives from puzzlement about how a mechanism (even a biological mechanism like the brain), if it proceeds in its lawful, mechanistic way, can ever produce novelty. The problem is at the center of Plato's central question in the *Meno*: how can an untutored slave boy be led through a geometric argument until he understands the proof? The answer Plato provides, which hardly satisfies our modern ears, is that the boy knew it all the time; his new understanding was simply a recollection of a prior understanding buried deep in his memory (a recognition or intuition?). What bothers us about the answer is that Plato does not explain where the buried knowledge came from.

[6] No one has offered an explanation of why unconscious work during interruption should be more effective for solution than the continuation of conscious work. The simplest hypothesis consistent with the data is that it isn't more effective.

Combinatorially Generated Novelty

Let's leave the *Meno* (I have offered a solution for the puzzle elsewhere[7], and in any event, we are talking science here, not philosophizing), and go directly to the question of how a mechanism creates novelty, for novelty is at the core of creativity. In fact, we shall define creativity operationally, in full accordance with general usage, as novelty that is regarded as having interest or value (economic, esthetic, moral, scientific or other value).

I shall start with an example. There are about 92 stable elements in nature, composed of protons and neutrons (and these, in turn, of component particles) There are innumerable molecules, chemical species, almost none of which existed just after the Big Bang or just after the 92 elements first appeared in the universe.

Here is novelty on a mind-boggling scale; how did it come about? The answer is "combinatorics." Novelty can be created, and is created, by combinations and recombinations of existing primitive components. The 26 letters of the alphabet (or, if you prefer the 70-odd phonemes of English) provide the primitives out of which a denumerable infinity of words can be created. New numbers, new words, new molecules, new species, new theorems, new ideas all can be generated without limit by recursion from small finite sets of primitives.

The traditional name in AI for this basic novelty-producing mechanism is *generate and test.* One uses a combinatorial process to generate new elements, then tests to see if they meet desired criteria. A good example of a generate-and-test system that can create novelty valuable for science is the BACON program (Langley, Simon, Bradshaw and Zytkow, 1987). BACON takes as inputs uninterpreted numerical data and, when successful, produces as outputs scientific laws (also uninterpreted) that fit the data .[8]

[7] Simon (1976). I'll offer just a hint here. Having a test for recognizing the solution to a problem if it is attained (which the slave boy has) does not provide a path (a proof) that leads step by step from the given information to the solution. Theorem and proof path are wholly independent objects.

[8] I hasten to add that BACON has discovered no new scientific laws (although other programs built on the same generate-and-test principle have); but it has *rediscovered*, starting with only the same data that the original discoverer had, a number of the most important laws of 18th and 19th Century physics and chemistry.

Selective Search as Inspiration

The law-generating process that BACON uses to find laws that describe data is not a random search process. The space of "possible functions" is not finite, and even if we limited search to some finite portion of it, any useful domain would be too large to yield often to random search. Basically, BACON's law generator embodies three heuristics for searching selectively: First, it starts with simple functions, then goes on (by combinatorial means) to more complex ones. We don't have to pause long to define "simple" or "complex." The simple functions are just those primitive functions that BACON starts with (in fact, the linear function); the compound functions are formed by multiplying or dividing pairs of functions by each other. A function is "simple" if it is generated early in the sequence, "complex" if generated later.

Second, BACON is guided by the data in choosing the next function to try. In particular, it notices if one variable increases or decreases monotonically with respect to another, testing whether ratios of the variables are invariant in the first case, products in the second, and shaping the next function it generates accordingly. This simple operation generates a wide class of algebraic functions, and by enlarging a bit the set of primitive functions (e.g., adding the exponential, logarithmic and sine functions), the class of generatable functions could be greatly broadened. The main point is that BACON's choice of the next function to test depends on what kind of fit with the data the previously tried functions exhibited.

Third, in problems involving data about more than two variables, BACON follows the venerable experimental procedure of changing one independent variable at a time. Having found conditional dependencies among small sets of variables, it explores the effects of altering other variables.

That is essentially all there is to it. With these simple means, and provided with the actual data that the original discoverers used, BACON rediscovers Kepler's Third Law (It finds $P = D^{3/2}$ on the third or fourth try), Ohm's Law of current and resistance, Black's Law of temperature equilibrium for mixtures of liquids and a great many others. There are many other laws it *doesn't* discover, which is an essential fact if it is to be regarded as a valid theory of human performance. Humans also *don't* discover laws more often than they discover them.

To validate BACON as a theory of human discovery, we would like to have as detailed historical data as possible on how the human discoveries were actually made, but sometimes the data are quite scanty. About all we know about Kepler's discovery of his Third Law is that he initially made a mistake, declaring that the period of revolution of the planets varied as the square of their distance from the Sun. Some years later, he decided the fit of law to data was poor and went on to find the correct law. Interestingly enough, BACON first arrives at Kepler's erroneous square law, rejects it as not fitting the data well enough, and goes on to the correct law almost immediately. With a looser parameter to test whether a law fits the data, BACON would make Kepler's mistake.

Sometimes the processes of BACON can be tested directly against human processes. Yulin Qin and I (1990) gave students the data (from the World Almanac) on the periods and distances of the planets — labeling the variables simply *x* and *y*, without interpretation. In less than an hour, 4 of 14 students found and fitted the 3/2-power law to the data. The students who succeeded used a function generator that responded to the nature of the misfits of the incorrect functions. The students who failed either were unable to generate more than linear functions or generated functions whose form was independent of previous fits and misfits.

I spell out this example to show that theories of inspiration are constructed and tested in exactly the same manner as other scientific theories. Once the phenomena have been defined, we can look for other phenomena that accompany them and for mechanisms that exhibit the same behavior in the same situations. In historical cases more favorable than Kepler's, we may have voluminous data on the steps toward discovery. In the case of both Faraday and Krebs, for example, laboratory notebooks are available, as well as the published articles and autobiographical accounts. In these cases, we have many data points for matching the scientist's behavior with the model's predictions.

Discovery of New Concepts

I have now cited a few pieces of evidence — many more exist — that scientists do not have to be "seized by the god" to discover new laws; such laws, even laws of first magnitude, can be arrived at by quite understandable and simulatable psychological processes. But what about new concepts? Where do they come from?

BACON is provided with one heuristic that I have not yet mentioned. When it discovers that there is an invariant relation in the interaction between two or more elements in a situation, it assigns a new property to the elements, measuring its magnitude by the relative strength of each element's action (one of the elements is assigned a unit value, becoming the standard). For example, BACON notices that when pairs of bodies collide, the ratio of accelerations of any given pair is always the same. BACON defines a new property (let's call it "obstinance"), and assigns an obstinance of 1 to body A, and an obstinance to each other body inversely proportional to the magnitude of its acceleration in collisions with A. Of course, *we* know that "obstinance" is what we usually call "inertial mass," and that BACON has reinvented that latter concept on the basis of this simple experiment.

This procedure turns out to be a quite general heuristic for discovering new concepts. BACON has used it to reinvent the concepts of specific heat, of refractive index, of voltage, of molecular weight and atomic weight (and to distinguish them) and others. Here again, inspiration turns out to be a by-product of ordinary heuristic search.

All of these results are available in the psychological and cognitive science literature (Langley, Simon, Bradshaw and Zytkow, 1987). They will not be improved by philosophical debate, but rather, by careful empirical study to determine the range of their validity and the goodness with which they approximate the observed phenomena. Debate, philosophical or otherwise, is pointless without familiarity with the evidence.

Other Dimensions of Discovery

Scientists do many things besides discovering laws and concepts. They plan and carry out experiments and interpret the findings, invent new instruments, find new problems, invent new problem representations. There are other dimensions to discovery, but these are perhaps the most important. I shall say no more about experiments (see Kulkarni and Simon, 1988) or instruments or problem-finding here. Some processes for finding new representations have already been examined in our discussion of insight. There is still plenty of work to be done, but so far, no evidence of which I am aware that the explanation of the phenomena of intuition, insight and inspiration will require the introduction of mechanisms or processes unlike those that have been widely employed in simulating human thinking. That, of course, is an empirical claim — actually, not so much a claim as an invitation to join in the exciting task of explaining how machines like people and computers can think, and sometimes think creatively.

Neurophysiological Foundations

It will not have passed without notice that I have said almost nothing today about the brain as a physiological organ. My silence should not be interpreted as doubt that the mind is in the brain, or a suggestion that processes beyond the physiological are required for its operation. The reason for my omission of the physiology of the brain is quite different. As I have pointed out in other contexts, sciences generally progress most effectively if they focus upon phenomena at particular levels in the scheme of things. Hunters of the quark do not, fortunately, need to have theories about molecules, or vice versa. The phenomena of nature arrange themselves in levels (Simon, 1981) and scientists specialize in explaining phenomena at each level (high energy physics, nuclear physics, analytic chemistry, biochemistry, molecular biology . . . . neurophysiology, symbolic information processing, and so on), and *then*, in showing (at least in principle) how the phenomena at each level can be explained (reduced) to the terms and mechanisms of the theory at the next level below.

At the present moment in cognitive science, our understanding of thinking at the information processing level has progressed far beyond our knowledge of the physiological mechanisms that implement the symbolic processes of thought. (Fortunately, on the computer side, we know full well how the symbolic processes are implemented by electronic processes in silicon.) Our ignorance of neurology is regrettable but not alarming for progress at the information-processing level, for this same sky-hook picture of science is visible in every scientific field during some period — usually a long period — in the course of its development. Nineteenth Century chemistry had little or no base in physics, and biology had only a little more in chemistry.

There is no reason why research in cognition should not continue to develop vigorously at both physiological and information processing levels (as it is now doing) watching carefully for the indications, of which there already are a few, that we can begin to build the links between them — starting perhaps with explanations of the nature of the physiological mechanisms (the "chips" and "integrated circuits") that constitute the basic repositories of symbolic memory in the brain. While we await this happy event, there is plenty of work for all of us, and no lack of knowledge of cognitive mechanisms at the symbolic level I have been considering in this paper.

Some Philosophical Implications

Several questions of major interest to philosophy that are closely connected with, cognition are empirical questions whose which cannot be solved by fact-free speculation, no matter how sophisticated it may be. A major difficulty with these questions is that finding empirical data to answer them appears to require us to look inside the human head, which is not easy to do, especially if introspection is ruled out as an incorrigibly solipsistic process. This presents a difficulty, but not an insuperable difficulty. The view that we cannot build testable theories of the processes within the head, including the processes of thought, is no more tenable than the view that biochemical theory cannot capture the laws of life. With the coming of computers and the demonstration that they can model not only the products but also the processes of thought, this mental vitalism is no longer defensible.

Testability of Theories of Mental Phenomena

To say that there are many variables within the head that are not directly observable is simply to say that a theory of mental phenomena will contain theoretical terms, not a novelty for any of the sciences. In such situations we need to insure that there is a sufficiently high ratio of observables to unknowns in our theories so that the values of the theoretical terms are overdetermined, hence ascertainable by convergent methods and testable. (Simon, 1970, 1983, 1985; Shen and Simon, 1993). When we construct a theory of mental phenomena in the form of a computer simulation, we test it by observing human subjects and a computer program performing exactly the same tasks, with identical inputs of stimuli. Then we compare the trace of the computer, at an appropriate level of detail, with observations of the human behavior (including verbal behavior) over the same interval of time. The examples provided in this paper have illustrated how this strategy has been employed to validate computer models of intuition, insight and creativity.

Specifically, to answer the question of whether an appropriately programmed computer can think, we establish a task and a set of criteria to determine whether a human being is thinking when performing that task. If the computer, given the same task, not only produces the same result but also matches the behavior of the human in all observable respects, and in particular, matches the processes the human is observably using during performance of the task, then we conclude that the computer is also thinking — i.e., that the processes that produce the result for the computer can be mapped on the processes that produced the same result for the human.

Thus, to determine whether the theoretical term "thinking" applies to the computer, we use the same test that we use to determine whether it applies to the human subject. Of course, we do not in this way find any magic that solves the problem of Hume — we do not *prove* that our theory of thinking is correct: but merely that it is compatible with the available empirical evidence. Again, this does not distinguish methods of theory verification in psychology from those in any other science. In no science does research *prove* the correctness of a theory; at best it shows that it has not been falsified and provides a reasonable fit to some body of facts.

The Mind-Body Problem

Suppose, now, that we have constructed a computer program that passes this test of thinking, for some range of tasks. We can now ask what solution, if any, it offers to the mind-body problem. It was Carnap, in 1955, who first explicitly proposed this use of the computer as a tool in epistemology.[9] This is the way he put his proposal:

> In order to make the method of structure analysis applicable, let us now consider the pragmatic investigation of the language of a robot rather than that of a human being. In this case we may assume that we possess much more detailed knowledge of the internal structure. . . . Just as the linguist [e.g., Quine's linguist in *Word and Object*], . . . begins with pointing to objects, but later, after having determined the interpretation of some words, asks questions formulated by these words, the investigator of [the robot's] language . . . begins with presenting objects . . . but later, on the basis of tentative results concerning the

[9] Carnap (1955) reprinted in Carnap (1956).

> intensions of some signs . . . proceeds to present predicate expressions . . . which use only those interpreted signs. . . .
>
> Instead of using this behavioristic method, the investigator may here use the method of structure analysis. On the basis of the given blueprint of [the robot], he may be able to calculate the responses which [it] would make to various possible inputs. In particular, he may be able to derive from the given blueprint . . . fairly precise boundaries for the intensions of certain concepts. . . .
>
> It is clear that the method of structure analysis, if applicable, is more powerful than the behavioristic method, because it can supply a general answer and, under favorable circumstances, even a complete answer to the question of the intension of a given predicate. . . .
>
> The intension of a predicate can be determined for a robot just as well as for a human speaker, and even more completely if the internal structure of the robot is sufficiently known to predict how it will function under various conditions.

With the advance of computers and programming languages in the years since Carnap made his proposal, we can now describe, in detail, computer programs that, by using a physical symbol system to carry out thought, embody a clear answer to the mind-body problem. Just as a brain uses neurons and associated tissues to store information (in ways that we do not understand in detail), so a computer uses physical devices (of quite diverse mechanical, electrical, and electronic varieties) to store information (in ways that we do understand in detail). What is required in both cases is a system built of components that can be maintained, with some stability, in one or another of two or more states, and that can input and output information by signaling the current states of these components. The specific substances of which these memories are built, and the physical or biological processes they use to maintain and alter memory contents are relevant only in fixing the capacity, stability, and speed of the system, and do not limit its basic qualitative capabilities.

Now we put human subjects and a computer program (the latter named GPS (Newell & Simon, 1972), or Soar (Newell, 1990), or Act (Anderson, 1983)) to work solving the Tower of Hanoi puzzle,[10] comparing the think-aloud protocols of the humans with the trace of the program. We find that both ultimately solve the puzzle, making certain characteristic errors along the way.

For example, if the puzzle has an odd number of disks, both computer and humans will often initially make the wrong first move; much less often if there are an even number of disks. The reason for this is that the correct (but counterintuitive) move, in the former case but not the latter, is to place the small disk on the target peg which will later need to be clear so that the largest disk can be moved to it. The mistake will be made by some versions of the computer program but not by others (and by some people and not others) depending on the way in which goals (intentions) are formed by the computer, and the degree of foresight (look-ahead) that is exercised in forming them. Examination of the human protocols reveals that the same phenomena of goal formation determine who does or doesn't make the error. (Anzai & Simon, 1979). Further exploration of the data shows that means-ends analysis ("find differences between current and goal states, find operators that usually remove such differences, apply operators," etc.) characterizes much of the behavior of both humans and the computer programs.

By means of these and other observations, we develop variants of the computer programs that match the differences in behavior among subjects, thereby obtaining an explanation not only of how "people" solve (or fail to solve) the Tower of Hanoi problem, but also what differences in strategy lead different people to follow different solution paths (Simon, 1975). The computer programs can also include learning mechanisms (e.g., adaptive production systems) that match, and explain, the changes in the strategy of individual human problem solvers as they acquire skill in solving the problem.

In what sense do these findings constitute a solution of the mind-body problem? They show that a demonstrably mechanistic system (a "body" in the form of a physical symbol system) is capable of thinking (using "mind" processes), where the operational definition of "thinking" is identical with the definition used to determine when people are thinking. It cannot be emphasized too strongly that the

[10] The Tower of Hanoi puzzle consists of a set of discs of different sizes impaled in pyramidal fashion on one of three pegs. The task is to move the disks to form a pyramid on another one of the pegs, moving only one disk at a time, and never placing a larger disk atop a smaller one.

operational test of thinking involves comparison of both product and process. With this definition and these empirical findings, the research in cognitive science has shown that a mind is simply a brain at work.

If we wish to preserve the two terms, "mind" and "brain," in our language, then we can use the former for the processes of thought, and the latter for the structure that supports the processes. There is no more mystery in the relation between these two components than in the relation of the cardboard of which an old IBM punchcard is fabricated, and the punching of a pattern of holes in it. The former is the memory, a part of the brain; the latter is the process of storing knowledge in the brain. To describe any dynamic system — the planets revolving about the Sun, or an electric generator — we must describe both the physical parts as organized, and the processes they undergo: organized substance and process. The brain and mind are a dynamic system; hence their description takes this same form. There is nothing epiphenomenal about mind, for without process, the brain does not think. As one component of the system is substance, the other process, they are not identical.

The Chinese Room

Searle has rejected this solution of the mind-body problem on the grounds that the wrong definition of thinking has been applied. Thinking, he argues, requires *understanding* the object of thought; and computers, he claims, cannot understand. He provides as example the parable of a room in which translation from English to Chinese (or vice versa) is going forward, but simply by means of a lexicon that finds the proper Chinese translation for each English word (or phrase, or sentence) without reference to the word's intension. Hence, the translation can be done without understanding of either language.

The answer to Searle is that he has described the wrong room. If the room had windows, so that the translators could see in the real world instantiations of the situations described by the text, then they (assumed to know English) could acquire Chinese meanings, building up their lexicon in the form of a huge discrimination net that sorts both situations and linguistic expressions according to their sensed properties (the intensions), and associates expressions with the situations they denote. Now the Chinese text is associated with its intensions, and these are used to associate to English text corresponding to these intensions. A computer system, ZBIE, which carries out these processes was constructed and described a quarter century ago by Siklóssy (1972). It could also be used to construct a Chinese-English lexicon (or vice versa), using the associations of both languages with their intensions to link the former.

Siklóssy's demonstration that a computer can learn the intensions of words, phrases and sentences shows empirically that computers are capable of thinking even if the "thinking" is so defined as to require knowledge of the intensions of the symbols the mind is manipulating in the process of thought. Hence, our solution of the mind-body problem remains valid even if we use this stricter definition of thinking and mind.

Conclusion

Artificial intelligence is an empirical science with two major branches. One branch is concerned with building computer programs (and sometimes robots) to perform tasks that are regarded as requiring intelligence when they are performed by human beings. The other is concerned with building computer programs that simulate, and thereby serve as theories of, the thought processes of human beings engaged in these same tasks. I have directed my remarks to the outer edge of AI research belonging to the latter branch, where it is concerned with phenomena that are often regarded as ineffable, and not explainable by machine models. I have shown that, on the contrary, we have already had substantial success in designing and implementing empirically tested information-processing theories that account for the phenomena of intuition, insight and inspiration. I have no immediate urge to predict how much further we shall go in the future or how fast. The continual progress on the journey over the past forty years has been speedy enough for me.

I have had some harsh things to say about philosophers and philosophy (perhaps no harsher than philosophers have had to say about AI). Of course I am not really attacking philosophers but rather those people who think they can reach an understanding of the mind and of the philosophical questions surrounding it by methods other than those of empirical psychological science. Traditional philosophy has much more to learn today from AI than AI has to learn from philosophy, for it is the human mind we must understand — and understand as a physical symbol system — in order to advance our understanding of the classical questions that philosophers have labeled "epistemology" and "ontology" and the "mind-body problem" (Simon, 1992).

My argument stands on a solid body of fact. I have mentioned a considerable number of these facts, drawn from papers in refereed journals or similarly credible sources. I may perhaps be pardoned for drawing a large portion of the facts I have cited from work in which I have been involved. I could have made an even stronger case if I had broadened the base, but I would have been familiar with fewer of the details. If you want to calibrate my base of evidence, you can multiply it by several orders of magnitude to take account of the work of all the other members of the AI and cognitive science communities who have been engaged in simulation of human thinking. In my account, I have tried not to talk about "future hopes of understanding or modeling human thinking," but to confine myself to documented, easily replicable, present realities about our present capabilities for modeling and thereby explaining human thinking, even thinking of those kinds that require the processes we admiringly label "intuitive," "insightful," and "inspired."

I have used the mind-body problem to illustrate how cognitive science, using computer simulation as a tool of theory, can bring light to bear on important epistemological problems. The conclusion reached from a large and consistent body of empirical evidence is that brain and mind are simply the essential substance and process that define any system, computer or human, capable of thinking.

If I have challenged some dimensions of human uniqueness, I hope I will not be thought scornful of human beings, or of our capacity to think. To explain a phenomenon is not to demean it. An astrophysical theory of the Big Bang or a three-dimensional chemical model of DNA do not lessen the fascination of the heavens at night or the beauty of the unfolding of a flower. Knowing how we think will not make us less admiring of good thinking. It may even make us better able to teach it.

References

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

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86,* 124-140.

Carnap, R. (1955). Meaning and synonymy in natural languages. In Carnap (1956).

Carnap, R. (1956). *Meaning and necessity* (2nd ed). Chicago, IL: University of Chicago Press.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal Reports as Data* (rev. ed.). Cambridge, MA: The MIT Press.

Feigenbaum, E. A. & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8,* 305-336.

Kaplan, C. A. (1989). *Hatching a theory of incubation.* Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA.

Kaplan, C. A. & Simon, H. A. (1990). In search of insight. *Cognitive Psychology, 22,* 374-419.

Kulkarni, D. & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12,* 139-176.

Langley, P., Simon, H.A., Bradshaw, G. L. & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes.* Cambridge, MA: The MIT Press.

McClelland, J. L. & Rumelhart, D. E. (1986). *Parallel distributed processing* (volumes 1 and 2). Cambridge, MA: The MIT Press.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

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

Plato, *The Meno*

Quillian, R. (1967). *Semantic memory.* Unpublished doctoral dissertation, Department of Psychology, Carnegie Institute of Technology.

Qin Y. & Simon, H. A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science, 14,* 281-312.

Richman, H. B., Staszewski, J. J. & Simon, H. A., (1995). Simulation of expert memory using EPAM IV. *Psychological Review, 102,* 305-330.

Shen, W. (1994). *Autonomous learning from the environment.* New York, NY: W. H. Freeman.

Shen, W., & Simon, H. A. (1993). Fitness requirements for scientific theories containing recursive theoretical terms. *British Journal for the Philosophy of Science, 44,* 641-652.

Siklóssy, L. (1972). Natural language learning by computer. In H. A. Simon & L. Siklóssy, (Eds.), *Representation and Meaning.* Englewood Cliffs, NJ: Prentice-Hall.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Childrens's thinking: What develops?* Hillsdale, NJ: Lawrence Erlbaum.

Simon, H. A. (1970). The axiomatization of physical theories. *Philosophy of Science, 37,* 16-26.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7,* 268-288.

Simon, H. A. (1983). Fitness Requirements for scientific theories. *British Journal for the Philosophy of Science, 34,* 355-365.

Simon, H. A. (1985). Quantification of theoretical terms and the falsifiability of theories. *British Journal for the Philosophy of Science, 36,* 291-298.

Simon, H. A. (1976). Bradie on Polanyi on the Meno paradox. *Philosophy of Science, 43,* 147-151.

Simon, H. A. (1979, 1989). *Models of Thought.* New Haven, CT: Yale University Press

Simon, H. A. (1996). *The sciences of the artificial,*(3rd ed.). Cambridge, MA: The MIT Press.

Simon, H. A. (1992). The computer as a laboratory for epistemology. In L. Burkholder (Ed.), *Philosophy and the computer.* Boulder, CO: The Westview Press.