Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration

Herbert A. Simon

Departments of Computer Science and Psychology
Carnegie Mellon University

Complex Information Processing
Working Paper #529

April 1995

Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration

Herbert A. Simon
Departments of Computer Science and Psychology
Carnegie Mellon University

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.

1 Introduction

I am deeply honored to receive this mark of esteem and friendship from my colleagues in artificial intelligence. It is now just forty years since Al Newell, Cliff Shaw and I took the plunge into the exhiliarating waters of AI. We called it "complex information processing," but that label had no chance of survival in competition with John McCarthy's more vivid "artificial intelligence," a term that he introduced at about the same time.

Whatever the label, research in AI has provided as exciting and satisfying a lifetime adventure as any scientist could desire. For me, the stored-program computers that first came to my attention about 1950 brought an end to many years of frustration, during which I had been searching for a formal language capable of expressing theories of human thinking, problem solving and decision making. Beginning in the middle 1950s, Al Newell, Cliff Shaw and I undertook to use computer languages for this purpose.

2 Computer Programs as Theories

During the 1930s and '40s, and into the early '50s, I had 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 any kind, changed all that by permitting one to implement a very literal representation of human symbol processing in the machine's memories and processes.

Computer programs written in these 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.[1]

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 LISP, and 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.

3 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 today. 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 languagesare tested in exactly the same way

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

as theories written in differential equations. We use the theories to make predictions, which are then tested against behavior captured in the laboratory or observed in the field.[2]

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 preceeding 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.

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.

Some of you may have heard me express these last sentiments before or have read them in my published papers (Newell and Simon, 1972). I feel obliged to repeat them 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. I must confess that I have not in recent years read any books of this kind, but I have seen reviews of them. Before you convict me of bigotry for ignoring them, let me explain why I do so.

3.1 Cognition'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.[3] 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), and especially if we wish to broadcast the results of our evaluations, 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. (Notice that I am not asking anyone to read it, but just to refrain from public comment if they haven't read it.)

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 and 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 successful 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 and Simon, 1984; and Richman, *et al*., 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 and 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 and Simon, 1988), LIVE (Shen, 1994) and others that simulate many of the discovery processes

[2] The theories of physics consist not only of the differential equations, but also deducible properties of these equations (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, in EPAM, the short-term memory capacity can be deduced from the structure and parameters of the program.

[3] I should perhaps explain that I selected the topic and title of my talk for this occasion before learning that there would be a session at this meeting on AI and philosophy -- news that I obviously greeted without enthusiasm.

that are discernable in the activities of scientists. Some of the models I have mentioned are complementary, some are competitive, as theories are in any science.

Again, I must remind you that 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, you 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.

3.2 Limits of Explanation?

In the remainder of my talk 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 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.

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 unable 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 process 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.

4 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 macros, the intermediate steps of which are these intuitions, the expert takes long strides in search, the novice tiny steps.

4.1 A Theory of Intuition

After specifying how we shall 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, 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 nearly 10^5 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 current interest is not in machine competence in medical diagnosis but in models of intuition. EPAM, as described, is exhibiting intuition, and modeling at least the first stage of thought (the recognition stage) of an experienced physician confronted with a set of symptoms.

4.2 Testing the Recognition Model

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).

[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.

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 and 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 the 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.

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.

5 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.

5.1 A Theory 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 atttentional 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.

5.2 Empirical Tests of the Theory

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 and 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 representration 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-by-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.

5.3 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 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 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 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.

6 Inspiration

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.

6.1 Generating 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

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

[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.

[7] Simon (1976).

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, Bradsaw and Zytkow, 1987). BACON takes as inputs uninterpreted numerical data and, when successful, produces as outpouts scientific laws (also uninterpreted) that fit the data .[8]

6.2 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 functions 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 attend 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.

6.3 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,

[8] I hasten to add that BACON has discovered no new scientific laws (although other programs built in 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.

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.

6.4 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.

7 Physiological 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 a 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 physiology 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.

8 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 constructing and implementating 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.

With the privilege that age carries, of being curmudgeonly, 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 a class of people called "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).

Moreover, it is not really the privilege of age I am claiming; rather, it is the privilege that comes from standing 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. Given the nature of this occasion, I may perhaps be pardoned for drawing a large portion of the facts I have cited from work in which I have been involved directly. 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. So if you want to calibrate my base of evidence, you can multiply it by a couple of 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 carefully 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."

If I have been scornful of (some) philosophers, 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

Ericsson, K. A. and Simon, H. A. (1993), *Protocol analysis: Verbal Reports as Data*, Revised edition, Cambridge, MA: The MIT Press.

Feigenbaum, E. A. and 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.* Unpubl. doctoral thesis, Dept. of Psychology, Carnegie Mellon University, Pittsburgh, PA.

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

Kulkarni, D. and 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. and Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes.* Cambridge, MA: The MIT Press.

McClelland, J. L. and 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. and Simon, H. A. (1972), *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall

Quillian, R. (1967). *Semantic memory.* Unpubli. doctoral thesis, Dept. of Psychology, Carnegie Institute of Technology.

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

Richman, H. B., Staszewski, J. J. and Simon, H. A., (1995), Simulation of expert memory using EPAM IV. *Psychological Review* (forthcoming, April).

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

Simon, D. P. and 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. (1976), Bradie on Polanyi on the Meno paradox. *Philosophy of Science,*

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

Simon, H. A. (1981), *The sciences of the srtificial,* Second edition. 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.