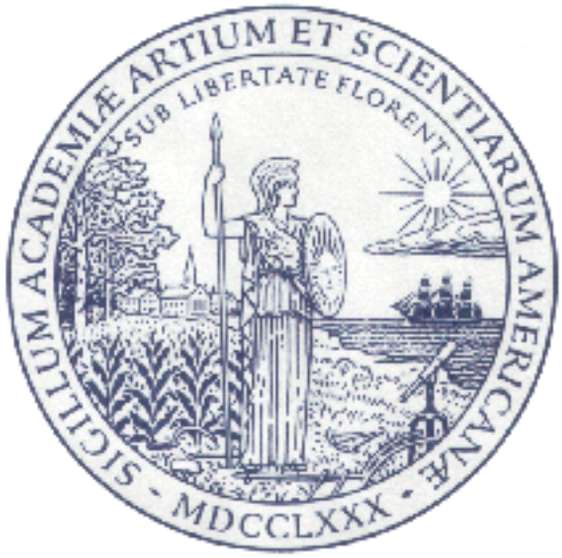

Stated Meeting Report

From Socrates to Expert Systems: The Limits of Calculative Rationality

Hubert L. Dreyfus

This year Artificial Intelligence (AI) is celebrating its thirtieth birthday; obviously an appropriate occasion for a retrospective evaluation.

AI began auspiciously, with Allen Newell and Herbert Simon's work at RAND. Newell and Simon proved that computers could do more than calculate. They demonstrated that computers were physical symbol systems whose symbols could be made to stand for anything, including features of the real world, and whose programs could be used as rules for relating these features. In this way computers could be used to simulate certain important aspects of intelligence. Thus the information-processing model of the mind was born. But this model of the mind as a symbol processor has run into trouble. Indeed, looking back over these thirty years, it seems that whereas practical AI is becoming more and more useful, theoretical AI appears more and more to be a perfect example of what Imre Lakatos has called a "degenerating research program."

Newell and Simon's early work on problem solving was impressive, and by 1970 Artificial Intelligence had turned into a flourishing research program, thanks to a series of micro-world successes such as Terry Winograd's SHRDLU, a program that could respond to English-like commands by moving simulated, idealized blocks. The field had its own Ph.D. programs, professional societies

---

References may be found in the original text of this communication. For a detailed treatment of the issues discussed, see Hubert Dreyfus and Stuart Dreyfus, *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, The Free Press, 1986.

and gurus. It looked like all one had to do was extend, combine, and render more realistic the micro-worlds and one would have genuine artificial intelligence. Marvin Minsky, head of the MIT program, predicted that "within a generation the problem of creating 'artificial intelligence' will be substantially solved."

Then, rather suddenly, the field ran into unexpected difficulties. The trouble started, as far as we can tell, with the failure of attempts to program children's story understanding. The programs lacked the common sense of a four-year-old. And no one knew what to do about it. An old philosophical dream was at the heart of the problem. AI is based on an idea which has been around in philosophy since Descartes, that all understanding consists in forming and using appropriate symbolic representations. For Descartes these were complex descriptions built up out of primitive ideas or elements. Kant added the important idea that all concepts were rules. Frege showed that rules could be formalized so that they could be manipulated without intuition or interpretation. Given the nature of computers, AI took up the search for formal rules and representations. Common-sense understanding had to be understood as some vast body of formalized propositions, beliefs, rules, facts and procedures. And it simply turned out to be much harder than one expected to formulate, let alone formalize, the required theory of common sense. It was not, as Minsky had hoped, just a question of cataloguing a few hundred thousand facts. The common-sense knowledge problem became the center of concern. Minsky's mood changed completely in the course of fifteen years. He told a reporter: "The AI problem is one of the hardest science has ever undertaken."

Given this impasse, it makes sense to return to micro-worlds—domains isolated from everyday common-sense under-

standing—and at least try to develop theories of such domains. This is actually happening—with the added realization that such isolated domains need not be games like chess nor micro-worlds like Winograd's blocks world but can be skill domains like disease diagnosis or spectrograph analysis.

Thus, from the frustrating field of AI has recently emerged a new field called knowledge engineering, which by limiting its goals has applied this research in ways that actually work in the real world. The result is the so-called expert system, which has been the subject of recent cover stories in *Business Week* and *Newsweek* and Edward Feigenbaum's book *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. According to a *Newsweek* headline: "Japan and the United States are rushing to produce a new generation of machines that can very nearly think."

Feigenbaum, one of the original developers of expert systems (who stands to profit greatly from this competition) spells out the goal:

> In the kind of intelligent system envisioned by the designers of the Fifth Generation, speed and processing power will be increased dramatically; but more important, the machines will have reasoning power: they will automatically engineer vast amounts of knowledge to serve whatever purpose humans propose, from medical diagnosis to product design, from management decisions to education.

What the knowledge engineers claim to have discovered is that in areas which are cut off from everyday common sense and social intercourse, all a machine needs in order to behave like an expert are some general rules and lots of very specific knowledge. This specialized knowledge is of two types:

> The first type is the *facts* of the domain—the widely shared knowledge . . . that is written in textbooks and journals of the

> field . . . . Equally important to the practice of the field is the second type of knowledge called *heuristic knowledge*, which is the knowledge of good practice and good judgment in a field . . . that a human expert acquires over years of work.

Using all three kinds of knowledge Feigenbaum developed a program called DENDRAL. It takes the data generated by a mass spectrograph and deduces from this data the molecular structure of the compound being analyzed. Another program, MYCIN, takes the results of blood tests such as the number of red cells, white cells, sugar in the blood, etc., and comes up with a diagnosis of which blood disease is responsible for this condition. It even gives an estimate of the reliability of its own diagnosis. In their narrow areas, such programs give impressive performances. They seem to confirm Leibniz's claim:

> [T]he most important observations and turns of skill in all sorts of trades and professions are as yet unwritten. This fact is proved by experience when, passing from theory to practice, we desire to accomplish something. *Of course, we can also write up this practice, since it is at bottom just another theory more complex and particular* . . . .

And, indeed, isn't the success of expert systems just what one would expect? If we agree with Feigenbaum that: "almost all the thinking that professionals do is done by reasoning . . . " we can see that once computers are used for reasoning and not just computation they should be as good or better than we are at following rules for deducing conclusions from a host of facts. So we would expect that if the rules which an expert has acquired from years of experience could be extracted and programmed, the resulting program would exhibit expertise. Again Feigenbaum puts the point very clearly:

> [T]he matters that set experts aside from beginners, are symbolic, inferential, and rooted in experiential knowledge .... Experts build up a repertory of working rules of thumb, or "heuristics," that, combined with book knowledge, make them expert practitioners.

Since each expert already has a repertory of rules in his mind, all the expert system builder need do is get the rules out and program them into a computer.

This view is not new. In fact, it goes back to the beginning of Western culture, when the first philosopher, Socrates, stalked around Athens looking for experts in order to draw out and test their rules. In one of his earliest dialogues, *The Euthyphro*, Plato tells us of such an encounter between Socrates and Euthyphro, a religious prophet and so an expert on pious behavior. Socrates asks Euthyphro to tell him how to recognize piety: "I want to know what is characteristic of piety . . . to use as a standard whereby to judge your actions and those of other men." But instead of revealing his piety-recognizing heuristic, Euthyphro does just what every expert does when cornered by Socrates. He gives him examples from his field of expertise, in this case mythical situations in the past in which men and gods have done things which everyone considers pious. Socrates persists throughout the dialogue in demanding that Euthyphro, then, tell him his rules for recognizing these cases as examples, but although Euthyphro claims he knows how to tell pious acts from impious ones, he cannot state the rules which generate his judgments. Socrates ran into the same problem with craftsmen, poets and even statesmen. None could articulate the principles underlying his expertise. Socrates concluded that none of these experts knew anything and he didn't either. Not a promising start for Western philosophy.

Plato admired Socrates and saw his problem. So he developed an account of

what caused the difficulty. Experts, at least in areas involving non-empirical knowledge such as morality and mathematics, had, in another life, learned the principles involved, Plato said, but they had forgotten them. The role of the philosopher was to help such moral and mathematical experts recollect the principles on which they act. Knowledge engineers would now say that the rules experts—even experts in empirical domains—use have been put in a part of their mental computers where they work automatically.

> When we learned how to tie our shoes, we had to think very hard about the steps involved .... Now that we've tied many shoes over our lifetime, that knowledge is "compiled," to use the computing term for it; it no longer needs our conscious attention.

On this Platonic view, the rules are there functioning in the expert's mind whether he is conscious of them or not. How else could one account for the fact that the expert can perform the task? So nothing has changed. Only now 2000 years later, thanks to Feigenbaum and his colleagues, we have a new name for what Socrates and Plato were doing: *knowledge acquisition research.*

But although philosophers and even the man in the street have become convinced that expertise is based on applying sophisticated heuristics to masses of facts, there are few available rules. As Feigenbaum explains: "[A]n expert's knowledge is often ill-specified or incomplete because the expert himself doesn't always know exactly what it is he knows about his domain." So the knowledge engineer has to help him recollect what he once knew:

> [An expert's] knowledge is currently acquired in a very painstaking way; individual computer scientists work with individual experts to explicate the expert's heuristics—to mine those jewels of knowl-

> edge out of their heads one by one .... [T]he problem of knowledge acquisition is the critical bottleneck in artificial intelligence.

When Feigenbaum suggests to an expert the rules the expert seems to be using, he gets a Euthyphro-like response: "That's true, but if you see enough patients/rocks/chip designs/instrument readings, you see that it isn't true after all," and Feigenbaum comments with Socratic annoyance: "At this point, knowledge threatens to become ten thousand special cases."

There are also other hints of trouble. Ever since the inception of Artificial Intelligence, researchers have been trying to produce artificial experts by programming the computer to follow the rules used by masters in various domains. Yet, although computers are faster and more accurate than people in applying rules, master-level performance has remained out of reach.

Arthur Samuel's work is typical. In 1947, when electronic computers were just being developed, Samuel, then at IBM, decided to write a checker playing program. He elicited heuristic rules from checker masters and programmed a computer to follow these rules.

The resulting checkers program is not only the first and one of the best experts ever built, but it is also a perfect example of the way fact turns into fiction in AI. Feigenbaum, for example, reports that "by 1961 [Samuel's program] played championship checkers, and it learned and improved with each game." In fact, Samuel said in a recent interview at Stanford University, where he is a retired professor, that the program did once defeat a state champion but the champion "turned around and defeated the program in six mail games." According to Samuel, after 35 years of effort, "the program is quite capable of beating any amateur player and can give better players a good contest." It is clearly no champion.

Samuel is still bringing in expert players for help but he "fears he may be reaching the point of diminishing returns." This does not lead him to question the view that the masters the program cannot beat are using heuristic rules; rather, like Plato and Feigenbaum, Samuel thinks that the experts are poor at recollecting their compiled heuristics: "The experts do not know enough about the mental processes involved in playing the game."

The same story is repeated in every area of expertise, even in areas unlike checkers where expertise requires the storage of large numbers of facts, which should give an advantage to the computer. In each area where there are experts with years of experience the computer can do better than the beginner, and can even exhibit useful competence, but it cannot rival the very experts whose facts and supposed heuristics it is processing with incredible speed and unerring accuracy.

In the face of this impasse, in spite of the authority and influence of Plato and 2000 years of philosophy, we must take a fresh look at what a skill is and what the expert acquires when he achieves expertise. We must be prepared to abandon the traditional view that a beginner starts with specific cases and, as he becomes more proficient, abstracts and interiorizes more and more sophisticated rules. It might turn out that skill acquisition moves in just the opposite direction: from abstract rules to particular cases. Since we are all experts in many areas, we have the necessary data, so let's look and see how adults learn new skills.

Stage 1: Novice

Normally, the instruction process begins with the instructor decomposing the task environment into context-free features which the beginner can recognize without benefit of experience. The beginner is then given rules for determining actions on the basis of

these features, like a computer following a program.

For purposes of illustration, let us consider two variations: a bodily or motor skill and an intellectual skill. The student automobile driver learns to recognize such interpretation-free features as speed (indicated by his speedometer) and distance (as estimated by a previously acquired skill). Safe following distances are defined in terms of speed; conditions that allow safe entry into traffic are defined in terms of speed and distance of oncoming traffic; timing of gear shifts is specified in terms of speed, etc. These rules ignore context. They do not refer to traffic density or anticipated stops.

The novice chess player learns a numerical value for each type of piece regardless of its position, and the rule: "Always exchange if the total value of pieces captured exceeds the value of pieces lost." He also learns that when no advantageous exchanges can be found, center control should be sought, and he is given a rule defining center squares and one for calculating extent of control. Most beginners are notoriously slow players, as they attempt to remember all these rules and their priorities.

Stage 2: Advanced Beginner

As the novice gains experience actually coping with real situations, he begins to note, or an instructor points out, perspicuous examples of meaningful additional components of the situation. After seeing a sufficient number of examples, the student learns to recognize them. Instructional maxims now can refer to these new *situational aspects* recognized on the basis of experience, as well as to the objectively defined *non-situational features* recognizable by the novice.

The advanced beginner driver uses (situational) engine sounds as well as (non-situational) speed in his gear-shifting rules. He shifts when the motor sounds like it is

straining. He learns to observe the demeanor as well as position and velocity of pedestrians or other drivers. He can, for example, distinguish the behavior of the distracted or drunken driver from that of the impatient but alert one. No number of words can take the place of a few choice examples in learning these distinctions. Engine sounds cannot be adequately captured by words, and no list of objective facts enables one to predict the behavior of a pedestrian in a crosswalk as well as can the driver who has observed many pedestrians crossing streets under a variety of conditions.

With experience, the chess beginner learns to recognize over-extended positions and how to avoid them. Similarly, he begins to recognize such situational aspects of positions as a weakened king's side or a strong pawn structure despite the lack of precise and universally valid definitional rules.

Stage 3: Competence

With increasing experience, the number of features and aspects to be taken account of becomes overwhelming. To cope with this information explosion, the performer learns, or is taught, to adopt a hierarchical view of decision-making. By first choosing a plan, goal or perspective which organizes the situation and by then examining only the small set of features and aspects that he has learned are relevant given that plan, the performer can simplify and improve his performance.

A competent driver beginning a trip decides, perhaps, that he is in a hurry. He then selects a route with attention to distance and time, ignores scenic beauty, and as he drives he chooses his maneuvers with little concern for passenger comfort or for courtesy. He follows more closely than normal, enters traffic more daringly, occasionally violates a law. He feels elated when decisions work out and no police car appears,

and shaken by near accidents and traffic tickets.

The Class A chess player, here classed as competent, may decide after studying a position that his opponent has weakened his king's defenses so that an attack against the king is a viable goal. If the attack is chosen, features involving weaknesses in his own position created by the attack are ignored as are losses of pieces inessential to the attack. Removal of pieces defending the enemy king becomes salient. Successful plans induce euphoria and mistakes are felt in the pit of the stomach.

In both of these cases, we find a common pattern: detached planning, conscious assessment of elements that are salient with respect to the plan, and analytical rule-guided choice of action, followed by an emotionally involved experience of the outcome.

The experience is emotional because choosing a plan, a goal or perspective is no simple matter for the competent performer. Nobody gives him any rules for how to choose a perspective, so he has to make up various rules which he then adopts or discards in various situations depending on how they work out. This procedure is frustrating, however, since each rule works on some occasions and fails on others, and no set of objective features and aspects correlates strongly with these successes and failures. Nonetheless the choice is unavoidable. While the advanced beginner can hold off using a particular situational aspect until a sufficient number of examples makes identification reliable, to perform competently requires choosing an organizing goal or perspective. Furthermore, the choice of perspective crucially affects behavior in a way that one particular aspect rarely does.

This combination of necessity and uncertainty introduces an important new type of relationship between the performer and his environment. The novice and the advanced

beginner, applying rules and maxims, feel little or no responsibility for the outcome of their acts. If they have made no mistakes, an unfortunate outcome is viewed as the result of inadequately specified elements or rules. The competent performer, on the other hand, after wrestling with the question of a choice of perspective or goal, feels responsible for, and thus emotionally involved in, the result of his choice. An outcome that is clearly successful is deeply satisfying and leaves a vivid memory of the situation encountered as seen from the goal or perspective finally chosen. Disasters, likewise, are not easily forgotten.

Remembered whole situations differ in one important respect from remembered aspects. The mental image of an aspect is flat; no parts stand out as salient. A whole situation, on the other hand, since it is the result of a chosen plan or perspective, has a "three-dimensional" quality. Certain elements stand out as more or less important with respect to the plan, while other irrelevant elements are forgotten. Moreover, the competent performer, gripped by the situation that his decision has produced, experiences the situation not only in terms of foreground and background elements but also in terms of opportunity, risk, expectation, threat, etc. As we shall soon see, if he stops reflecting on problematic situations as a detached observer, and stops thinking of himself as a computer following better and better rules, these gripping, holistic experiences become the basis of the competent performer's next advance in skill.

Stage 4: Proficiency

Considerable experience at the level of competency sets the stage for yet further skill enhancement. Having experienced many situations, chosen plans in each, and having obtained vivid, involved demonstrations of

the adequacy or inadequacy of the plan, the performer involved in the world of the skill, "notices," or "is struck by" a certain plan, goal or perspective. No longer is the spell of involvement broken by detached conscious planning.

Since there are generally far fewer "ways of seeing" than "ways of acting," after understanding without conscious effort what is going on, the proficient performer will still have to think about what to do. During this thinking, elements that present themselves as salient are assessed and combined by rule to produce decisions about how best to manipulate the environment.

On the basis of prior experience, a proficient driver approaching a curve on a rainy day may sense that he is traveling too fast. Then, on the basis of such salient elements as visibility, angle of road bank, criticalness of time, etc., he decides whether to take his foot off the gas or to step on the brake. (These factors would be used by the *competent* driver to *decide that* he is speeding.)

The proficient chess player, who is classed a master, can recognize a large repertoire of types of positions. Recognizing almost immediately and without conscious effort the sense of a position, he sets about calculating the move that best achieves his goal. He may, for example, know that he should attack, but he must deliberate about how best to do so.

Stage 5: Expertise

The proficient performer, immersed in the world of his skillful activity, *sees* what needs to be done, but *decides* how to do it. With enough experience with a variety of situations, all seen from the same perspective but requiring different tactical decisions, the proficient performer gradually decomposes this class of situations into subclasses, each of which share the same deci-

sion, single action, or tactic. This allows the immediate intuitive response to each situation which is characteristic of expertise.

The expert chess player, classed as an international master or grandmaster, in most situations experiences a compelling sense of the issue and the best move. Excellent chess players can play at the rate of 5–10 seconds a move and even faster without any serious degradation in performance. At this speed they must depend almost entirely on intuition and hardly at all on analysis and comparison of alternatives. My brother, Stuart, recently performed an experiment in which an international master, Julio Kaplan, was required rapidly to add numbers presented to him audibly at the rate of about one number per second, while at the same time playing five-second-a-move chess against a slightly weaker, but master level player. Even with his analytical mind completely occupied by adding numbers, Kaplan more than held his own against the master in a series of games. Deprived of the time necessary to see problems or construct plans, Kaplan still produced fluid and coordinated play.

Kaplan's performance seems somewhat less amazing when one realizes that a chess position is as meaningful, interesting and important to a professional chess player as a face in a receiving line is to a professional politician. Almost anyone can add numbers and simultaneously recognize and respond to faces, even though each face will never exactly match the same face seen previously, and politicians can recognize thousands of faces, just as Julio Kaplan can recognize thousands of chess positions similar to ones previously encountered. The number of classes of discriminable situations, built up on the basis of experience, must be immense. It has been estimated that a master chess player can distinguish roughly 50,000 types of positions.

Automobile driving probably involves the

ability to discriminate a similar number of typical situations. The expert driver, generally without any awareness, not only knows by feel and familiarity when an action such as slowing down is required, but he knows how to perform the action without calculating and comparing alternatives. He shifts gears when appropriate with no awareness of his acts. What must be done, simply is done.

It seems that a beginner makes inferences using rules and facts just like a heuristically programmed computer, but that with talent and a great deal of involved experience the beginner develops into an expert who intuitively sees what to do without applying rules. The tradition has given an accurate description of the beginner and of the expert facing an unfamiliar situation, but normally an expert does not *reason*. He does not *solve problems*. He does what normally works and, of course, it normally works.

Given this account of the five stages of skill acquisition, we can understand why the common-sense knowledge problem has proved to be so hard. Common-sense *understanding* might well be everyday *know-how*. By know-how I do not mean propositional knowledge nor even procedural rules, but knowing what to do in a vast number of special cases.

Common-sense physics, for example, has turned out to be extremely hard to spell out in a set of facts and rules. When one tries, one either requires more common sense to understand the facts and rules one finds or else one produces formulas of such complexity that it seems highly unlikely they are in a child's mind. It just may be that the problem of finding a *theory* of common sense in physics is insoluble. By playing with all sorts of liquids and solids for several years, the child may simply have developed an ability to discriminate thousands of typical cases of solids, liquids, etc., each paired with a typical skilled response to its typical behavior

in typical circumstances. There may be no theory of common-sense physics more simple than a list of all such typical cases and even such a list is useless without a similarity-recognition ability.

Our phenomenology of skill acquisition also enables us to understand why the knowledge engineers from Socrates, to Samuel, to Feigenbaum have had such trouble getting the expert to articulate the rules he is using. The expert is simply not following any rules! He is doing just what Feigenbaum feared he might be doing—discriminating thousands of special cases. This in turn explains why expert systems are never as good as experts. If one asks an expert for rules, one will, in effect, force the expert to regress to the level of a beginner and state the rules he still remembers but no longer uses. If one programs these rules on a computer, one can use the speed and accuracy of the computer and its ability to store and access millions of facts to outdo a human beginner using the same rules. But no amount of rules and facts can capture the knowledge an expert has when he has stored his experience of the actual outcomes of tens of thousands of situations.

The knowledge engineer might still say that in spite of appearances the mind and brain *must* be reasoning—making millions of rapid and accurate inferences like a computer. After all, the brain is not "wonder tissue" and how else could it work? But there *are* other models for what might be going on in the hardware that make no use of the sort of symbols Newell and Simon have in mind. That is, they do not use symbols that correspond to recognizable features of the world and rules that represent these features' relationships.

Researchers who call themselves "new connectionists," are building devices and writing programs that operate somewhat like neural nets. These parallel distributed processing systems can recognize patterns

and detect similarity and regularity without using inferences or isolated features at all. In a connectionist machine, the states of the machine cannot be interpreted as symbols representing invariant features of the skill domain. These connections models offer new hope for the success of AI once the field gives up the Newell/Simon hypothesis that to produce intelligence, computers must be used as physical symbol systems. Thanks to connectionism, computers may someday exhibit skill, and AI researchers may someday solve—or better, by-pass—the common-sense knowledge problem they have inherited from philosophy.

Once one gives up the assumption that intuitive experts must be making inferences, and admits the role of involvement and intuition in the acquisition and application of skills, one will have no reason to cling to the heuristic program as a model of human intellectual operations. Feigenbaum's claim that "we have the opportunity at this moment to do a new version of Diderot's *Encyclopedia*, a gathering up of all knowledge—not just the academic kind, but the informal, experiential, heuristic kind"; as well as his boast that thanks to Knowledge Information Processing Systems we will soon have "access to machine intelligence—faster, deeper, better than human intelligence" can both be seen as a late stage of Socratic thinking, with no rational or empirical basis. In this light those who claim we must begin a crash program to compete with the Japanese Fifth Generation Intelligent Computers can be seen to be false prophets blinded by Socratic assumptions and personal ambition—while Euthyphro, the expert on piety who kept giving Socrates examples instead of rules, turns out to have been a true prophet after all.

Hubert L. Dreyfus is Professor of Philosophy at the University of California at Berkeley. *His communication was presented at the 1671st Stated Meeting, held in Cambridge on October 8, 1986.*