JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1985, 44, 271-277 NUMBER 2 (SEPTEMBER)

THE KISS AND THE PROMISE: A REVIEW OF HUBERT L. DREYFUS' WHAT COMPUTERS CAN'T DO: THE LIMITS OF ARTIFICIAL INTELLIGENCE[1]

EDWARD K. CROSSMAN

UTAH STATE UNIVERSITY

In *What Computers Can't Do: The Limits of Artificial Intelligence,* philosopher Hubert L. Dreyfus provides a series of well developed arguments against the popular notion that a machine (digital computer) can exhibit behavior that approximates intelligent human behavior. At the outset, this book offers a much-needed introduction to the field of artificial intelligence (AI) and, relatedly, to cognitive simulation. This feature alone is sufficient recommendation, but there is more. The clarity of Dreyfus' writing and the care taken to lead the reader by the hand through the AI field with its technical jargon and concepts ensures that even a reader unfamiliar with AI will enjoy the book and will be able to explore further the fascinating world of machine-behavior analogies.

Many authors both inside and outside the field of AI extol its virtues and promises (e.g., Feigenbaum & Feldman, 1963; Feigenbaum & McCorduck, 1983; Jastrow, 1982; Minsky, 1966, 1968). The visible critics are few; in fact, only three names ordinarily surface: Weizenbaum, Dreyfus, and Searle. Joseph Weizenbaum (1976), who in 1966 developed the psychotherapeutic program ELIZA to simulate an interview between therapist and client, became appalled when people took the application seriously. His criticisms of AI center mostly upon the misuse of computers in society, and particularly upon the supplanting of human emotional interactions by machine-human interchange. Dreyfus' concerns are different; he attacks the AI enterprise directly by undercutting the very assumptions on which it rests. Both he and another philosopher, John Searle, known for his metaphor of the Chinese-speaking room (Searle, 1980), argue that the study of artificial intelligence with the digital computer sheds no light whatsoever on "meaning" or "understanding." They assert that the digital computer in principle is capable only of accepting input, translating that input according to a dictionary or a fixed set of rules, and outputting the translated information.

It is important, perhaps vitally so, for behavior analysts to gain familiarity with AI in general and with Dreyfus' arguments in particular. The enterprise predicated upon AI is rapidly gaining a substantial share of public attention and of both public and private resources. AI is rapidly evolving beyond the tinkerings of basement hobbyists who constructed crude robotic devices, to the major efforts of corporations to simulate human behavior in a variety of forms. A race is on, not only in the United States but also in France, England, and Germany. Japan, with its fifth-generation (intelligent-like) computers, has declared its interest in winning this race. Feigenbaum and McCorduck (1983) have argued that supremacy in the AI field is central to the economic growth and development of any nation in the 21st century. Behavior analysts cannot afford to ignore this trend.

In addition, the proponents of AI have issued a direct challenge to psychology in general and to behavior analysis in particular. Early on, this challenge took the form of ignoring psychology with its slow organic models. More recently, as Dreyfus discusses, pro-

[1]Dreyfus, H. L. (1979). *What Computers Can't Do: The Limits of Artificial Intelligence* (rev. ed.). New York: Harper & Row. xiii + 354 pp., including index.

Reprints may be obtained from Edward K. Crossman, Department of Psychology, Utah State University, Logan, Utah 84322-2810.

ponents of AI have begun to recognize cognitive and Gestalt approaches of psychology, because of the apparent similarities between these systems and concepts that are central to AI, such as memory, retrieval, and form perception.

Behaviorism is still viewed by AIers as largely irrelevant. Furthermore, behaviorism is condemned for suppressing the study of the "Mind" and its operations. These are familiar criticisms, but it should be noted that in this case they come from a viewpoint that shares the philosophical heritage of behaviorism. Aristotle, Plato, Descartes, the French materialists, and the British, French, and German empiricists provide the essential background for both approaches, and both behaviorism and AI characterize the organism in deterministic terms.

There are, however, several differences worth noting. Unlike behaviorism, AI has borrowed from idealists such as Kant the notion that there are innate rules that help organize incoming sensory data. This places AI within the dualist camp where the emphasis is on analyzing mediational processes rather than on objective behavior. Behavior analysts use computers to perform detailed and accurate analyses of relationships between environmental change and behavior change. In contrast, AI research, particularly early on, attempted to study behavior and to some extent the environment as a computerized representation embodied in a structure of formal rules. These differences placed behavior analysis and AI in opposing camps. Nevertheless, as AI begins to design systems that "learn" and takes a more functional approach, behavior analysts and AI workers will have more to talk about.

Meanwhile, it is advisable for behavior analysts to develop an understanding of the AI field. Dreyfus, although not a behavior analyst, offers an excellent starting point. His book provides both a survey of AI and a series of carefully crafted criticisms that, after some translation, most behavior analysts should find congenial. Dreyfus organizes his book along three main lines. The first part consists of four phases and is a chronological survey of major developmental periods in AI spanning the years 1957-1977. In the second part Dreyfus covers four assumptions—biological, psychological, epistemological, and ontological—that he believes underlie AI. In the final part he considers various alternatives to these assumptions.

THE FOUR PHASES

During the first phase (1957-1962), AI workers were "kissed" with a few successes, such as the Logic Theorist devised by Newell, Shaw, and Simon, which successfully proved 38 of 52 theories from *Principia Mathematica*; and the General Problem Solver (GPS) by Newell and Simon, which solved several complex routing problems. Prompted by these early successes, AI workers made a number of bold promises that were unkept; Dreyfus considers three. First, it was promised that automatic machine translation of various languages was close at hand. Despite large amounts of money dedicated to this project, the problems of semantics and syntactics in natural language overwhelmed the computers and their programs. Even today, the goal of automatic machine translation seems as elusive as in the early sixties. The promises in two other areas, problem solving and pattern recognition, also turned out to be empty. Recently, however, limited but notable progress has been made in problem solving and cognitive simulation, matters that will be discussed later.

Dreyfus labels the second phase of AI research (1962-1967) as the Semantic Information Processing period, after the title of a book by Minsky. During this phase the objective was to create a computer simulation of English language understanding, with "understanding" narrowly defined to mean that with a limited subset of English as input, the computer would respond with a limited subset of appropriate English. Thus, after the computer was told that it was on, when the user asked, "are you on?," the computer would output, "yes." Attempts to expand the subset of English met with failure, but AI workers continued to be optimistic. Proponents of AI such as Minsky explained away such failures in terms of technological limitations such as machine storage capacity.

The third phase (1967-1972) was characterized by microworlds, which is to say, by AI programs with greatly restricted domains. Terry Winograd's program SHRDLU is prototypical of AI work in this period. A human, using simple English commands, could cause a simulated robot arm to manipulate a set of variously shaped blocks. Although AI workers considered SHRDLU an advance in natural language understanding, Dreyfus views it as an admission of the fact that such programs can work only if the domain is artificially constrained. Behavior analysts may disagree with Dreyfus on this point, because the early steps taken by any science occur in a greatly simplified context. During this period, Dreyfus contends, AI's contributions to psychology, formerly touted by AI workers, also began to be questioned by psychologists such as Eleanor Rosch (1973; see also Dreyfus, 1979, p. 23), whose research on perception suggested a holographic model rather than the information-processing model upon which AI rested.

The retrenchment that occurred in the third phase was carried over into the fourth phase (1972-1977), which Dreyfus refers to as the Knowledge Representation or Cognitive Science phase. During this phase, "expert systems," the most recent development in AI, had their origin. For example, MYCIN, a program for diagnosing blood and meningitis infections, did a respectable job of simulating the expertise of a medical practitioner. Attempts to work in a broader context, however, where the relevance of facts could not be so precisely predetermined as in the MYCIN example, ran into trouble.

By this point if not before, the behavior analyst can certainly appreciate the difficulties that current AI workers face. For example, when an AI program or an expert system is forced to operate in a broader context than that for which it was designed, the problem becomes one of how to generalize and how to form the proper conditional discriminations. Dreyfus offers as an example, "The box was in the pen" (p. 215). This sequence has one meaning if uttered in a child's nursery but quite another meaning if uttered in a spy movie. These are only two of the theoretically infinite number of stimulus contexts the AI system would have to evaluate if it were always to respond appropriately to the statement. Not only would a search for the appropriate context take a very long time (arguably, a merely technological restriction); the question also remains as to what rules would be used to identify the proper context. Given that the von Neumann type of digital computer is strictly a formalistic, logic-following machine, Skinner's (1969) distinction between an organism's rule-governed and contingency-shaped behavior becomes critical. According to Skinner, some human behavior can be established on the basis of rule-following or instructional control. For example, the apprentice blacksmith may operate the bellows according to a memorized rule: "Up high, down low, up quick, down slow—and that's the way to blow" (Skinner, 1969, p. 139). The initial statement of the rule presumably grew out of a history of reinforcement contingencies; moreover, the selection of an effective rule given a particular situation and adherence to that rule are both instances of behavior presumably under the control of differential reinforcement. The problem, and it is central, is that it is not at all clear how the functioning of a computer can be reinforced. Without the ability to establish a history of differential reinforcement, many things become impossible, such as being able to invoke the correct rule in a situation where an infinite, or at least a very large, number of contextual stimuli are possible.

Although Dreyfus does not explicitly discuss the difference between contingency-shaped and rule-governed behavior, he does cite Polanyi's example of the bicycle rider who, in attempting to maintain balance, might be following the rule: "wind along a series of curves, the curvature of which is inversely proportional to the square of the velocity" (p. 190). Although such a rule may accurately describe what the bicycle rider is doing, Dreyfus is quick to point out that it does not provide an explanation of the behavior itself. Unfortunately, at this point Dreyfus suggests that the place to look for such explanations is in the brain or mind rather than in the history of reinforcement and punishment that the bicycle

rider has experienced. Still, the point stands that the computer rules that result in human-like behavior are similar to the physical laws that describe balance on a two-wheeled vehicle. Such rules may be useful as descriptions of behavior but not as explanations, and this is a blow to any cognitive science that maintains that a knowledge of such rules will lead to explanations of behavior. Although the relation between rule-governed and contingency-shaped behavior calls for more research, it is at least clear that behavior produced under instructional control has some properties different from behavior shaped by the relevant contingencies but in the absence of verbal descriptions of them (e.g., Matthews, Shimoff, Catania, & Sagvolden, 1977).

THE FOUR ASSUMPTIONS

According to Dreyfus, AI workers have been basing their efforts upon four assumptions he finds largely incorrect. Even though AI is rapidly changing and Dreyfus initially discussed these same assumptions in the 1972 edition of the book, most are as relevant today as when he first described them.

Biological assumption. This is the assumption that the brain and the digital computer are functionally similar—that each processes information both digitally (via on-off states) and serially. However, virtually no scientist is willing to describe the brain either as a primarily sequential device or as a digital one. In a series of articles comparing the operations of a human brain with those of the computer, psychologist Ernest Kent (1978a, 1978b, 1978c, 1978d) characterized both similarities and differences between the two. Basically, the brain's logic "gate," the neuron, sums inputs analogically and emits a stream of digital pulses through the axon. Many neurons behave in this manner simultaneously, thus differing from the successiveness of the rigid sequential stages that the digital computer must follow. But contemporary AI workers are not as bothered by these differences as they once were. Instead, they stress that the outputs of both systems must be similar, regardless of the type of intermediate processing.

Psychological assumption. Here cognitive scientists assert that even though brain-computer analogies may be weak, humans process discrete bits of information as does a computer. Dreyfus argues, however, that there is no evidence that humans search lists, sort, or classify neutral bits of information as a computer does. Instead, humans appear to follow a two-step process whereby raw visual and auditory stimuli are first translated into integrated wholes, such as visual images, melodies, etc. Next, humans manipulate these integrated wholes, as in comparing one song to another. Dreyfus here clearly favors a Gestalt interpretation and is referring to the problem of fuzzy sets, or stimulus classes whose boundaries are loosely defined.

Herrnstein, Loveland, and Cable (1976) also appreciated this characteristic of perception. They studied natural concepts by training pigeons to discriminate pictures of trees, of bodies of water, or of a particular person. Then, when presented with novel slides containing new instances of these objects, the birds were able to discriminate such slides from other slides not containing them. Accounting for the bird's classification in terms of attributes of the various stimuli proved to be an impossible task. Herrnstein et al. suggested the insufficiency of a theory of common elements, an approach contemporary AI might take. Both Dreyfus and Herrnstein et al. instead favor holistic theories, such as that of Eleanor Rosch (1973), which state that humans form a prototypical image of an average category member against which other stimuli are matched. The dimensions of such a prototype of course remain ill-defined, but the point according to Dreyfus is that present-day computers do not operate in this manner, and probably cannot do so without the participation of a human who discriminates the relevant stimulus class boundaries. If this is the case, there is little hope that a computer can adequately simulate psychological process.

Epistemological assumption. Even though humans may function in ways substantially different from the information-processing mode of the computer, there is still the possibility that human behavior can be formalized ac-

cording to another set of rules that can be reproduced by a machine. This is the epistemological assumption. For example, Dreyfus points out that planets traveling in their orbits are not solving differential equations. By specifying these equations, however, it is possible to construct artificial planets that behave as do the real planets. Might not human behavior also be formalized in the same sense? After all, humans are simply material beings subjects to the same physical laws and principles as other objects in the universe.

Unlike the psychological assumption, the epistemological assumption does not claim that the rules the computer follows in its simulation of human behavior provide any understanding of the causes of behavior. Perhaps the behavior analyst can feel more comfortable with the epistemological assumption because it avoids reference to the nature of mental processes, but Dreyfus cannot. Drawing from examples in natural language processing, Dreyfus persistently maintains that a rule-governed system is far too inflexible when it comes to interpreting statements that break the syntactic rules, as much discourse does. For example, the phrase "Rain slick careful" would confound the computer but not the human. Although it may be possible to give the computer rules for handling bad grammer, Dreyfus argues that even so, the human can come up with an exception to the rule and that this exception disproves that the machine is capable of human-like intelligent behavior.

Another related problem is metarules. The computer must be given rules that determine in a particular instance which other rules are applicable. But what determines which metarule is to be used? This leads to an infinite regress, according to Dreyfus a basically insoluble problem for the computer but not for the human.

Ontological assumption. This assumption, perhaps the most fundamental of all, maintains that the world can be exhaustively analyzed into context-free atomistic data or facts. How many facts? Perhaps 100,000 determinate facts would be enough to simulate intelligent behavior, but that problem is relatively small compared to the one of how to classify such a data base (Stevens, 1985). The computer, because it does not exist in a context, cannot determine independently of the human programmer what the appropriate context is for the moment. As a result, it cannot determine which facts are relevant and therefore should be manipulated. If the computer cannot discriminate between relevant and irrelevant facts, increases in memory size or processing speeds will not help.

Behavior analysts will again recognize this as a problem involving conditional discrimination. Some time ago, Goldiamond (1962) pointed out the importance of constant stimulus conditions that are not explicitly discriminative stimuli but that are inevitably coexistent with these stimuli and can alter their effects. These stimuli have been variously labeled "setting events," "establishing operations," and the like (Leigland, 1984); they are, in other words, contextual stimuli. A dirty joke told in church is likely to evoke a response completely different from that evoked by the same joke told at home. If an AI system can incorporate such contextual stimuli in its network of IF-THEN statements, perhaps this objection will dissolve. Dreyfus argues that the manner of contextual stimuli is unworkably large and hence poses an impossible task for present-day computers. Practically speaking, however, Dreyfus may be overstating his case. There are many instances where contextual stimuli can be specified, at least in limited domains, thus making it possible for the development of expert systems that will simulate human expertise. One such example is PROSPECTOR, a geological program that purportedly found the metal molybdenum on Mount Tolman where human geologists had failed (but see Dreyfus & Dreyfus, in press).

ALTERNATIVE ASSUMPTIONS

Dreyfus considers several other assumptions. Perhaps the most interesting of these is that intelligent human behavior may be simulated even though the computer has no body. If a human-like body, with its ability to sense the environment and to respond accordingly, is critical for intelligent behavior—and Drey-

fus believes that it is—then the computer is inadequate. To be sure, the sensory and locomotor capacities of the computer were not well developed at the time Dreyfus wrote. More recently, progress has been made, particularly in the areas of visual and voice recognition. These appear to be engineering problems that eventually will be solved.

Of greater interest to the behavior analyst, however, is the issue that, as Dreyfus expresses it, without a body the computer has no bodily "needs." In other words, there is no sense in which a computer can be deprived and hence no way in which the behavior of the computer can be altered through the application of reinforcers and punishers. It is not difficult for the behavior analyst to extend Dreyfus' language to make the point that the computer lacks the ability to be affected, as are humans, by environmental contingencies. Here the distinction between feedback and reinforcement is critical. The robotic computer equipped with limbs and a TV camera, having detected the presence of an object in the environment, say a wood block, can reach out and grasp the block or, based on feedback from its tactile sensors, can adjust limb position until the limb is oriented so that the fingers surround and grasp the block successfully. But if "grasping the block successfully" is to be considered as reinforcement of the chain of reaching and grasping responses, then a number of things should be altered. Obviously the probability of limb-extension and grasping in the presence of the block should increase, and it should decrease in its absence. Given identical stimulus situations and adjustment of the parameters of the feedback loops, errors can be eliminated totally, and the correct sequence of responses will occur all of the time. But when the behavior of a biological organism is similarly reinforced, not only is the probability of a highly specific response altered but so also are the probabilities of the members of an entire class of responses. The human, having learned to reach out and grasp a block, can also grasp a glass, a bottle, a book, and so on, even with spatial arrangements of such objects different from those when the initial reinforced response occurred.

This flexibility, or ability to generalize, is not a characteristic of modern-day von Neumann-type digital computers. Consistent with Herrnstein et al. (1976), determining the boundaries of a stimulus class following reinforcement in the presence of certain members of that class is an incredibly complex problem. Perhaps computers constructed around inorganic chips will never be able to exhibit the flexibility of contingency-shaped human behavior. Maybe the highly experimental field of "wet" engineering, in which organic chips are used to construct a molecular computer ("Working Toward," 1984), will ultimately offer a more promising direction for those who wish to construct a machine that will simulate a wider spectrum of human-like behavior.

CONCLUSION

Despite the criticisms raised by Dreyfus regarding the assumptions and over-inflated statements of accomplishment by those who work in the field of AI, there is much room to grow, and grow it will. Particularly, the demand for so-called "knowledge engineers" and the AI expert systems they devise will expand at an accelerating rate in the near future. That a strictly rule-governed system can ever simulate the richness of human behavior seems unlikely. Still, there are potential benefits to be derived from the AI field that the behavior analyst might enjoy. It seems reasonable to expect that the expansion of AI will focus greater public attention on behavior and on the necessity for understanding its environmental determinants. More specifically, as the AI field develops, its progress will surely be impeded at many different points. The behavior analyst who is aware of these impediments might be able to intervene, not only by providing information that originated in behavior-analytic research, but also by becoming involved with the techniques and procedures developed by engineers and computer scientists who are attempting to synthesize human-like behavior from an entirely different perspective. In so doing, the behavior analyst has the opportunity to transform the adversarial climate that now exists into a relation that will benefit all parties.

REFERENCES

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence* (rev. ed.). New York: Harper & Row.

Dreyfus, H. L., & Dreyfus, S. (in press). *Mind over machine: The power of human intuition and expertise in the era of the computer.* Riverside, NJ: Macmillan/The Free Press.

Feigenbaum, E. A., & Feldman, J. (Eds.). (1963). *Computers and thought.* New York: McGraw-Hill.

Feigenbaum, E. A., & McCorduck, P. (1983). *The fifth generation: Artificial intelligence and Japan's computer challenge to the world.* Reading, MA: Addison-Wesley.

Goldiamond, I. (1962). Perception. In A. J. Bachrach (Ed.), *Experimental foundations of clinical psychology* (pp. 280-340). New York: Basic Books.

Herrnstein, R. J., Loveland, D. H., & Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes,* **2,** 285-302.

Jastrow, R. (1982, June). The thinking computer. *Science Digest,* pp. 54-55, 106-107.

Kent, E. W. (1978a, January). The brains of men and machines. Part 1: Biological models for robotics. *Byte,* pp. 11-22, 96-106.

Kent, E. W. (1978b, February). The brains of men and machines. Part 2: How the brain controls outputs. *Byte,* pp. 84-90, 146-158.

Kent, E. W. (1978c, March). The brains of men and machines. Part 3: How the brain analyzes input. *Byte,* pp. 74-83, 94-108.

Kent, E. W. (1978d, April). The brains of men and machines. Part 4: The machinery of emotion and choice. *Byte,* pp. 66-89.

Leigland, S. (1984). On "setting events" and related concepts. *Behavior Analyst,* **7,** 41-45.

Matthews, B. A., Shimoff, E., Catania, A. C., & Sagvolden, T. (1977). Uninstructed human responding: Sensitivity to ratio and interval contingencies. *Journal of the Experimental Analysis of Behavior,* **27,** 453-467.

Minsky, M. L. (1966, September). Artificial intelligence. *Scientific American,* pp. 247-260.

Minsky, M. L. (Ed.). (1968). *Semantic information processing.* Cambridge, MA: MIT Press.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology,* **4,** 328-350.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences,* **3,** 417-457.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis.* New York: Appleton-Century-Crofts.

Stevens, L. (1985). *Artificial intelligence.* Hasbrouck Heights, NJ: Hayden.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* San Francisco: Freeman.

Working toward a molecular computer. (1984, September). *Chemical Week,* pp. 19-20.