BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 15, Number 2, October 1986

COMPUTERS AND THE NATURE OF MAN: A HISTORIAN'S PERSPECTIVE ON CONTROVERSIES ABOUT ARTIFICIAL INTELLIGENCE

BY JUDITH V. GRABINER

Throughout history, developments in the sciences have caused people to change their views of man and his place in the universe. The Copernican Revolution placed man on a planet, adrift in space; the Darwinian Revolution changed our view of human origins. Computers, too, raise questions about the nature of man: can computers think the way we do, and, if so, are we—like them—just thinking machines? Of course the question "can a machine think?" has been raised before, as long ago as the seventeenth century: by Descartes and Pascal, who said no; by Hobbes, who said that thought was mechanical; and somewhat later by La Mettrie, who saw man himself as a machine. But only in our century, as a result of work in the mathematical sciences, has the question "can a machine think?" been given widespread and rigorous discussion. Today computer scientists have devised programs that solve problems which, if solved by people, would seem to require intelligent thought. This point is made explicit by John McCarthy's name for the field: "artificial intelligence." As we shall see, the real controversy about artificial intelligence (AI) is not about the nature of computers and programs; it is about the nature of man.

But first, how would we possibly decide whether computers can think at all? Some 35 years ago, Alan Turing suggested a way of testing the assertion that a machine could think, without having to worry about the internal workings of the machine. If, he said, a machine could successfully imitate a human being in a full range of possible conversations, fooling its human conversational partner into believing the machine to be human, we ought to conclude that the machine was indeed thinking [**39**]. A machine that could do this would be said, in modern terminology, to have "passed the Turing test." Turing's proposed test has set the agenda for many subsequent debates over machine and human intellectual performance.

The purpose of the present paper is to provide a historical perspective on recent controversies, from Turing's time on, about artificial intelligence, and to make clear that these are in fact controversies about the nature of man. First, I shall briefly review three recent controversies about artificial intelligence, controversies over whether computers can think and over whether people are

---

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 01A60, 68-03; Secondary 68T01.

Based on a paper given at the AMS-MAA Joint Session on the History and Development of Modern Mathematics, Eugene, Oregon, August 17, 1984; received by the editors October 30, 1985.

0273-0979/86 $1.00 + $.25 per page

no more than information-processing machines. These three controversies were each initiated by philosophers who, irrespective of what the programs of their time actually did, viewed with alarm the argument that if a machine can think, a thinking being is just a machine. I will then turn to the major business of this paper: to contrast two developments from within the field of AI which have been interpreted by some as successful steps toward simulating human thought, and also to contrast some reactions to that claimed success. Finally, we will look at some recent developments in the field of AI that suggest that the whole discussion about machine intelligence is at best premature and at worst irrelevant.

Alan Turing himself expected that machines eventually would be able to pass his "imitation" test. One sophisticated objection, which he argued against, might be raised. Kurt Gödel had proved that there were limitations to the power of computers: any reasonably rich formal system is incomplete, and the consistency of such a system cannot be proved within the system. Turing in 1937 had himself shown that formal systems are equivalent to the behavior of machines. But Turing's reply was that people, just like computers, may well be subject to the limitations that Gödel had established [**39**, pp. 2109–2110]. Thus, Turing extrapolated from the "machine equals man" analogy to propose a limitation on human thought.

In 1961, the British philosopher J. R. Lucas, repelled by the idea that people are just instances of formal systems, refused to accept Turing's response to the Gödel-based objection. Lucas wanted to refute once and for all what he called "mechanism"—the view that the whole mind is just the sum of the operation of its separate parts. Lucas conceded that Gödel's incompleteness theorem applies to the machine. But, he argued, by standing outside the consistent, incomplete formal system, *we* can see some unprovable, meaningful formula to be true. The machine cannot produce the formula; we see the formula as true; so a human can beat the machine. Moreover, if the mind is just a formal system, Gödel's consistency theorem would say that the mind could not conclude itself to be consistent. But, Lucas continues, we do in fact assert our own consistency. Thus, no mechanical model of the mind can be adequate [**24**, pp. 115, 124].

Lucas's attempt to show that Gödel's theorem refuted mechanism provoked impassioned responses. (For a fuller account of Lucas's view and the responses to it, see [**16**].) Most important for our present purposes is that some of his critics accused Lucas of holding too exalted a view of man. For instance, if a specific machine cannot assert the true-though-unprovable Gödel formula, people cannot always do this either [**41, 44**]. Further, rather than suggest that our self-knowledge shows we are not machines, one can reverse the argument and say that, since formal systems cannot know themselves, Lucas's argument really implies that human self-knowledge is impossible [**2**, p. 30]. Again the statement "machine equals man" is used to imply that man has all the limitations of machines.

This controversy of the early 1960s made little reference to specific work on computer programs. But the 1960s saw the development of programs to play chess, to simulate human problem solving, and to carry out (at first in limited

areas) natural-language conversations. As often happens in an expanding new scientific field, some AI workers extrapolated their early successes, saying not "we have a chess program" or "we have a program that solves a class of puzzles like the 'cannibals-and-missionaries' problem," but "we're on the track of simulating thought." Thus the philosophical debate about what could be done, and about the nature of human thought, became bound up with the empirical questions of what had been done and of whether existing programs were intelligent. In response to the predictions of AI enthusiasts, philosopher Hubert Dreyfus said it couldn't be done, first in a paper provocatively entitled "Alchemy and AI" in 1965, and then in his book *What computers can't do*, first published in 1972. In the book, Dreyfus argued that not even the most highly touted AI programs of the time should be called "intelligent," and gleefully pointed out that the predictions for the field made by practitioners in 1960, "in ten years a computer will be the world's chess champion, ... discover and prove an important new mathematical theorem," and translate natural languages, had not yet come to pass [**11**, pp. 81–82, 91–92]. These things had not happened, Dreyfus argued, because they could not. AI research, he said, rests on the false assumption that the human mind works by operating on bits of information and performs its operations according to formal rules. Not so, said Dreyfus. Computers, in contrast to people, lack the ability to distinguish between the essential and the inessential [**11**, pp. **107, 112**]. Man is not "a fact or a set of facts," said Dreyfus, but "a being who creates both himself and the world of facts in the process of living in the world" [**11**, pp. 190–191]. If man is like this, programmed computers will never satisfy the Turing test. The partial successes of AI research are doomed to remain partial. Trying to imitate human intelligence with computers, in Dreyfus's view, is like trying to get to the moon by climbing trees [**27**, p. 180].

Practitioners in the field responded that Dreyfus's extrapolations about the limited future of AI research were the same kind of impossibility-argument as those of the seventeenth-century philosophers who denied that there could be a vacuum [**45**, p. 177]. For us, though, the key objection to Dreyfus will be about man, not programs. Dreyfus's alternate model of human thought was based not on neurophysiology or psychology, but on the philosophical school known as phenomenology, a position, according to his critics, neither empirically based nor intellectually rigorous. Dreyfus's point that we do not really understand how people think, so we cannot model the process with computers, was turned against him by Yorick Wilks, who said that we were so far from understanding the processes people use in thinking that the only way we even know that other human beings think is by the Turing test [**45**, p. 183].

Attacks on Dreyfus's argument that computers can never pass the Turing test make especially interesting the controversy initiated in 1980 by John Searle, who argued that even if a computer passed the test, it would not necessarily understand [**33**, pp. 417–419]. Searle's celebrated argument begins by imagining a person, knowing no Chinese, locked in a room with a lot of boxes filled with Chinese ideographs and with a book of rules in English that tells how to match up one Chinese ideograph with another. People outside the room pass the man Chinese symbols; obeying the instructions in the book, he

passes other Chinese symbols back out to them. His thoughts about the process are just, "Oh, it's squiggle, so I must give back squoggle." But the rules in the book are so thorough that the man's response to the symbols given him are what one would expect from a fluent speaker of Chinese. Thus, the man passes the Turing test, but he does not understand Chinese. Therefore, any machine of which he and his "program" are an instance does not understand Chinese either. Accordingly, Searle concludes that the Turing test is inadequate and that the thesis that a computer with an appropriate program is really a mind, or the analogous thesis that the mind is just an appropriately programmed computer, is wrong. The Turing test, said Searle, is "unashamedly behavioristic and operationalistic." People who accept its adequacy miss the distinction between simulation and duplication [**33**, p. 423].

One might think at first that Searle's point is irrefutable. After all, the man knows perfectly well that he does not understand Chinese. But the key responses to Searle's argument—that the total system (which includes the book of rules) understands Chinese, that rejecting behavioral equivalences and models would doom all science—were strong, and produced a robust debate [**16**, pp. 475–476; cp *Behavioral and Brain Sciences* **3**, 1980, pp. 417–457]. However, for our present task, the major point is that the purpose of Searle's Chinese-room thought-experiment was to reaffirm the complexity of human thought and the primacy of human intention. What he objected to was not the idea that collections of metals and circuits could learn and think, but to equating human understanding with merely being a formal system [**34**, p. 453].

All three of the controversies just described are about this question: if some kind of machine can do what a person can, is a person that kind of machine? But all these discussions have concerned the realm of philosophical possibility. Let us now pursue the discussion in the realm of scientific practice; let us examine some actual programs, the claims made about them, and some reactions to those claims.

In 1966, Joseph Weizenbaum of MIT described a computer program which could simulate, in limited circumstances, a natural-language conversation, playing an English-speaking psychotherapist of the school of Carl Rogers. This program, which Weizenbaum named ELIZA, impressed people tremendously. Here are some excerpts from Weizenbaum's published example of the program in action:

(Patient): Men are all alike.
(DOCTOR): IN WHAT WAY
(Patient): They're always bugging us about something or other.
(DOCTOR): CAN YOU THINK OF A SPECIFIC EXAMPLE
(Patient): Well, my boyfriend made me come here.

(DOCTOR): WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
(Patient): Perhaps I could learn to get along with my mother.
(DOCTOR): TELL ME MORE ABOUT YOUR FAMILY

(Patient): My father is afraid of everybody.
(DOCTOR): WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
(Patient): Bullies.
(DOCTOR): DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

[**42**, pp. 36–37][**43**, pp. 3–4].

This accomplishment—especially the DOCTOR's last question!—looks impressive, but is less so when the mechanism is revealed. The program has a script. It responds to specified cues in specified ways: for instance, to words like "mother" with "tell me more about your family"; to statements involving terms like "always" or "all" with requests for an example; and, most spectacularly here, to statements that do not fit the predetermined script by fishing with "does that have anything to do with" some previous topic whose importance to the patient was signalled by having been prefaced by the word "my."

Weizenbaum thought he had written a clever program, which did some interesting pattern matching. He didn't dream at first that people would believe that the program could actually think, much less act as though it were a caring therapist. But people did. It was suggested by some that the program had passed the Turing test [**16**, p. 476]. It was also suggested that such programs could practice therapy cheaply and on a large scale [**43**, p. 5; **9**, **31**]. ([**47**, p. 37] suggested recently that Weizenbaum's *intention* was to show how a computer could act like a psychologist.) But, knowing so well how little intelligence was involved in the program's plausible performance, Weizenbaum was shocked by responses like these. He was shocked because he thought these responses implied such widespread acceptance of the view that human beings were basically just complicated machines. In his 1976 book, *Computer power and human reason*, Weizenbaum characterized the information-processing model of man as just one aspect of a common twentieth-century mind-set that views human beings as means, rather than as ends, which falsely regards human problems as having technical solutions, and which encourages people to eschew moral choices in favor of "just following orders" [**43**, pp. 11, 13-14, 251, 275]. His book is an argument that the information-processing model of man is empirically false, and, even more important, that it is morally wrong. Like Dreyfus, Weizenbaum argued that people can do things machines cannot —for instance, we can understand natural language in a context of experiences like love and trust that machines cannot share [**43**, pp. 208–209]. Weizenbaum characterized even the most successful AI programs of the 1970s as lacking strong theory-based models of human intelligence, being instead collections of ad hoc programming tricks[1] [**43**, p. 232]. Weizenbaum concluded his critique with a call to those involved in the computer science profession "to teach the

[1] Weizenbaum develops at some length the argument that a programmer can, through trial and error, make some process "perform" a task even without the programmer's thoroughly understanding how it does it, in his ninth chapter, "Incomprehensible Programs." See, e.g., [**43**, pp. 234–236].

limitations of their tools as well as their power" [**43**, p. 277], and not to promulgate a view of human beings which helps to further dehumanize them. "What," he asks, "could it mean to speak of risk, courage, trust, endurance, and overcoming when one speaks of machines?" [**43**, p. 280].

Weizenbaum's book, like Dreyfus's, provoked outrage, and, what is more, attacks on Weizenbaum's scientific competence [**27**, p. 318]. His views on the potentially harmful directions in some AI research caused him to be confused with romantic critics of science and technology like Theodore Roszak, and provoked responses to his critique of AI which are in fact general defenses of the benevolence of technology. But Weizenbaum grants the value to mankind of science in general and computer science in particular. In any case, these responses do not vitiate Weizenbaum's central point: thinking of people as programmed machines will affect the decisions we make about how to treat people. To understand the nature of man, Weizenbaum defends the legitimacy of philosophy, literature, and emotion—which he calls "softer ways of knowing," concluding, "The computer is a powerful new metaphor for helping us to understand many aspects of the world, but... it enslaves the mind that has no other metaphors and few other resources to call on" [**43**, p. 277].

We now turn from Weizenbaum's work to a development in the field of AI from the 1980s. Gary Bradshaw, Pat Langley, and Herbert Simon have devised a family of computer programs called BACON (after the seventeenth-century philosopher Francis Bacon, who promoted inductive logic) that make scientific discoveries [**36, 4, 23, 35**]. That is, given data, these programs discover "a set of generalizations, or theory, to describe the data parsimoniously or to explain them" [**36**, p. 7]. This is not just a matter of fitting curves to points on a graph. The program generates hypotheses according to a set of heuristics, tests the hypotheses against data, and when necessary generates new theoretical terms to formulate the relevant law. The programs are not limited to one subject; they have been tested on a range of examples including Ohm's law, the ideal gas law, Snell's law of refraction, Kepler's third law of planetary motion, and Joseph Black's law of temperature equilibrium of mixtures [**4, 23, 35, 36**]. In particular, in discovering Black's law, the program invents a new theoretical term—one which Black, too, invented, and which we now call specific heat[2] [**4**,

[2] Described at greatest length in [**4**, pp. 972–973], but also described elsewhere, e.g. [**36**, p. 19; **23**, p. 125; **35**, 257–258]. In [4], Bradshaw, Langley, and Simon take Black's law to predict the final temperature of the mixture of two liquids and state it as follows. Consider substances #1 and #2, with masses m_1 and m_2, specific heats c_1 and c_2, and initial temperatures T_1 and T_2. Let the final temperature after they are mixed be f. Then Black's law states:

$$c_1 m_1 T_1 + c_2 m_2 T_2 = (c_1 m_1 + c_2 m_2) f.$$

The data available to the program in discovering the law, then, includes three classes of variables: some way of designating the substances involved (say by the subscripts 1 and 2), their respective masses, and the corresponding initial temperatures. As the authors point out, "the equation *can* be interpreted as saying that the total heat is conserved" [**4**, p. 973, my italics]. The authors then sketch a path followable by the computer, a path "that does not differ in essential respects from those that were tested." This "path" to the discovery I will simply quote [**4**, p. 973]:

> Initially, we employ identical volumes of the same substance (for example, water). In the first experiment, we vary the temperature of the first component, holding all other variables constant. We discover that the equilibrium

pp. 972–973]. (See also [**23, 35, 36**].) This work has attracted widespread attention in the scientific community far beyond AI circles.[3]

A program which makes scientific discoveries would indeed be a remarkable achievement, and its makers are well aware of the potential importance of what they are trying to do. They have made two major historical claims: first, that their program comes close to doing what Snell or Kepler or Black actually did; second, that their simulation helps us to understand some apparent philosophical difficulties raised by the real history of scientific discoveries. These claims are sufficiently important to quote exactly: "We confront the program with discovery problems that scientists have encountered, and we observe whether the program can make the discovery, *starting from the same point the scientists did*" (my italics) [**4**, p. 971]. (This is a sort of Turing test for scientific-discovery programs.) And, "The BACON experiments help to explain... how

---

> temperature is a linear function of the independent variable, with a slope of $\frac{1}{2}$. In the second experiment, we vary the second temperature, using symmetry between the two volumes to conjecture that the equilibrium temperature will also vary linearly with this new independent variable, with a slope of $\frac{1}{2}$. The data confirm this relation.
>
> In the third experiment, we vary the mass of the first component to determine how that mass enters the function determining the equilibrium. We now conjecture that a symmetric function will describe the joint effects of both masses. The data confirm this conjecture.
>
> In the fifth experiment, we change the composition of the first component (for example substitute mercury for water). BACON finds that a new coefficient must be introduced into the equation, whose value changes with change in the composition of the component. In the sixth experiment, we change the composition of the second component, and, by symmetry, introduce another new coefficient into the equation. The equation, thus modified, again fits the data. The coefficients introduced in the fifth and sixth experiments are, of course, the quantities that Black called specific heats.

The authors state that c_1 and c_2 are "new intrinsic properties" discovered by the program [**4**, p. 972]. (They are, of course, what we call the specific heats of the two substances; the term Black introduced was "capacity for heat." [**30**, p. 24] quotes Black's introduction of the latter term, after which Black added, "if I may be allowed to use this expression.")

As the authors claim, their search for the appropriate law for the given data is guided by a number of fruitful heuristics: that "extensive quantities" will be conserved [**35**, p. 258]; that the first relations to be looked at should be mathematically simple; that "similar variables will enter into laws symmetrically" [**35**, p. 258]; that "if two quantities covary (countervary), test their ratio (product) for invariance" [**35**, p. 254]; etc. These are, indeed, among the heuristics hard-won by centuries of work in the exact sciences. But they do not include heuristics for identifying the initial set of variables—a crucial task whose historical importance we shall discuss below.

[3] Witness the publication of their work in *Science* [**4**], Simon's Gibbs lecture at the American Mathematical Society's St. Louis meeting, January 1984, subsequently published in the Bulletin of the American Mathematical Society [**35**]. See also the report from a psychologist that "a computer program, given access to data available to past scientists, was able to rediscover such things as Ohm's law, Galileo's law of falling bodies, and one of Kepler's laws of planetary motion.... By all indications, this progress will continue and computers may soon far surpass human capability" [**13**, p. 143]. Psychologist Paul E. Meehl, in his foreword to [**13**], warns us not to be shocked by "the notion of using a computer to invent theories" just because the idea "is deeply threatening to the scientist's self-image." The possibility, he says, "is music of the future" [**13**, xxii].

some of the important laws of modern science [e.g., Kepler's third law] could antecede by decades or even generations the theoretical structures that subsequently rationalized them and took them out of the category of brute empirical generalizations." (As they point out, Kepler's laws long preceded Newton [**4**, p. 974].)

To these historical claims, Bradshaw, Langley, and Simon add a psychological one: that the computer model is adequate, not only for ordinary thought, but even for the thought of creative scientists. They grant that scientific discovery differs from ordinary problem solving in some ways: in being social, not individual, and in having indefinite goals rather than prespecified ones. Nonetheless, they maintain that there is "no reason to believe that the discoveries of human scientists cannot in time be explained within the information processing paradigm for problem solving" [**36**, p. 26]. We are certainly back at the heart of AI's claim to simulate how people really think. Bradshaw, Langley, and Simon base their confidence in this psychological prediction on the present accomplishments of their program. Have they now done what they claim? That is, does the BACON program in fact simulate real scientific discoveries like those of Black, Snell, or Kepler? I think not. To evaluate the claim, let us look at what is known about these discoveries.

Joseph Black, unlike the program, was in part guided to his discovery by what we now think of as a false theory of heat—that heat behaved like a substance which was conserved.[4] Even more important, Black's discovery was not merely a matter of applying this idea about heat to data about masses and temperatures of mixtures—the data the program uses for its "discovery." The essential difference between Black's discovery and the program's is that the program had its variables chosen for it. Black did not. In fact, before Black the very distinction between heat and temperature was not understood.[5] Furthermore, Black had to examine a wide range of phenomena, including experiments of his own and experiments reported in the literature.[6] These experiments were not conveniently labelled as to which would be important for a

---

[4] One evidence of this is Black's term "capacity for heat" rather than our modern "specific heat." The most readable account of Black's discovery in its historical context, along with lengthy quotations from the original sources, is that in [**30**]. The most relevant work of Black's predecessors is described in [**30**, pp. 20–29], and Black's own discussion is excerpted in [**30**, pp. 20–26]. Especially instructive is Black's discussion of theories of heat [**30**, pp. 42–45]. Black's work is also extracted in [**26**, pp. 134–139], a source Bradshaw, Langley, and Simon cite in their own bibliography in [**4**]. The classic historical account is that in [**28**], and good discussions may also be found in [**18**] and [**20**, pp. 75–79, especially p. 78]. These authors differ about the degree and date of Black's commitment to the theory of heat as a material substance, but all agree about his commitment to the idea that heat is conserved.

[5] The clear distinction between heat and temperature does not precede, but arises from, Black's work; see [**30**, pp. 17–18], showing where Black built on the work of Francis Bacon and G. Fahrenheit.

[6] Black culled key experiments from the writings of men like Herman Boerhaave, Pieter v. Musschenbroeck, George Martine, and others [**30**, pp. 23–28]; compare [**28**, pp. 13–15]. Heilbron [**20**, pp. 75–78] and Guerlac [**18**, pp. 178–179] point out in addition that Black seems to have discovered latent heat first. Thus, his experiments on specific heat were conditioned by that earlier work, and by the precise measurements of heat of fusion—premised in turn on the conservation of heat—that accompanied it.

yet-undiscovered law. In fact, Black had to disentangle those in the literature from theory-laden descriptions by authors whose views he did not fully share.[7] Moreover, a number of well-known experiments involving the transfer of heat had to be *neglected* by Black. For instance, the mixture could not allow changes of state.[8] Again, blacksmiths beat an iron rod over and over until the rod becomes hot enough to light fires. Black could not explain this experiment, but he dismissed it anyway as not sufficiently general to refute the theory of conservation of heat, pointing out that such an increase in temperature does not occur when one bangs equally hard on a similar rod of very elastic steel.[9] Thus, Black was guided to the specific experiments that shed light on the phenomena by theory. His discovery was not made by looking for regularities in a given set of data. Once one has chosen the appropriate data, the job of discovering Black's law is essentially over. Thus, the claim of BACON's creators that "the initial conditions for the simulation" are "what was already known at the time the discovery was made"[10] [**4**, p. 971] is false.

What of Snell's seventeenth-century law of refraction? Snell's law was "discovered" by the BACON program from data on the ratios of the path lengths of light beams and the perpendicular distances to the interface between the media, thus, the authors claim, avoiding any need for the program to know trigonometry [**23**, p. 124]. This is in fact perfectly fair; the law was stated by Snell himself (though the authors seem unaware of this) using lines and their ratios (though not the *same* lines), without explicit reference to sines [**38, 10**, pp. 389–390]. But even in this relatively simple example, some key conceptual

---

[7]Boerhaave, for instance, had argued (as had Musschenbroeck) that the "substance" heat was uniformly distributed in bodies at the same temperature. This, Black observed, was contradicted by the experiment of Fahrenheit, as reported by Boerhaave himself, of the result of mixing mercury and water at different temperatures; in this experiment, it was necessary to mix "three measures [*volumes*, by the way, not masses] of quicksilver with two of water, in order to produce the same middle temperature that is produced by mixing equal measures of hot and cold water" [**26**, p. 136; **30**, p. 23]. Boerhaave's theory is explained at length, with full attention to its Newtonian origins, in [**7**, pp. 214–234, esp. pp. 230–232]; a good, brief account, including its importance for Black's discovery, may be found in [**20**, pp. 61–63, 78]. The key point is that heat, for Boerhaave (he calls it "fire"), is a nondestructible substance made of particles [**7**, pp. 227, 230]. Only from the standpoint that heat is something that is conserved is Fahrenheit's experiment anomalous in the light of Boerhaave's theory; thus, Black's recognition of the conservation of heat is essential to his recognizing that Fahrenheit's experiment is important at all.

[8]As we have already pointed out, Black knew the law governing changes of state; he measured what he called the latent heat required to melt ice, giving the value of 143 Btu/lb [**30**, p. 37] in a paper read at Glasgow in 1762.

[9]Roller [**30**, p. 43] cites this experiment with iron, and the counterexample of steel [**30**, p. 44], from Black's works, pp. 33–34.

[10][**4**, p. 971] Occasionally Bradshaw, Langley, and Simon appear to moderate their claims in specific cases, suggesting a role for theory. But the role they describe does not seem to exceed suggesting that certain variables represent "extensive magnitudes" which are conserved, for a nonetheless predetermined class of what they call the experimental or empirical data, e.g. [**4**, p. 974; **35**, pp. 250, 257–258]. This does not do justice to the role of theory in Black's work in selecting data and experiments, and even defining what heat and temperature meant, as we have discussed above. The passages just cited from [**4** and **35**] are outweighed by the repeated claims of the authors that they are modeling the historical processes of significant scientific discoveries [**35**, pp. 260–261; **4**, especially title and abstract, p. 971; **23**, p. 25]; etc.

steps have been made for the program. Refraction had been studied for hundreds of years before Snell, but scientists of the caliber of Ptolemy thought that the angles, not the lines and their ratios, were the relevant variables; Ptolemy found no law relating them [**8**, pp. 271–281]. Moreover, in the seventeenth century Johann Kepler tried unsuccessfully to find a trigonometric relation between the angle of incidence and the angle of refraction. Unfortunately, he used as the "angle of refraction" the angle between the refracted ray and the (prolonged) incident ray. He didn't find Snell's law either [**3**, pp. 186–187]. Of course, Snell's law is not expressible as an elementary function even of the "correct" angles. Thus, the hard thing, historically, was looking for the ratio of the sines and identifying the relevant angles. Once we know to look for trigonometric functions of these particular angles, or for ratios involving precisely these lines, the job of discovering Snell's law is essentially done.

Let us now examine Bradshaw, Langley, and Simon's second historical claim: that they can explain how, in the history of science, a discovery can long antedate the theory that makes sense of it. We will use their example of Kepler's third law, which they call "a product of pure Baconian induction" [**4**, p. 973]. The relevant numerical variables, they say, are a planet's distance from the sun, D, and the period of its orbit, P. The program then discovers that D^3 varies as P^2, which is Kepler's law [**36**, pp. 13–14]. They say that these numerical variables are "observables"[11] [**36**, p. 13]. But they are not. What is observable in nature are planetary positions in the sky. Kepler had some very accurate ones, recorded by Tycho Brahe. But Kepler in 1619 had something else—a controversial theory, published in 1543 by Copernicus, that the planets went around the *sun*, and not, as was widely believed, around the earth. Given this theory and those observations an accomplished mathematician like Kepler could calculate the orbits, conclude that they were ellipses (Kepler's first law, 1609), and then determine the planetary periods and distances relative to the sun. Kepler also believed that a force from the sun was needed to push the heavy planets around, a force that diminished as one got farther from the sun, so that the speeds of the planets might well depend on their distance from the sun.[12] Then, guided by his strong belief in mathematical harmonies [**6**, pp. 138–139]—a belief shared by the BACON program—Kepler looked for the relationship between period and distance, and discovered his third law. Kepler's discovery was not a "brute empirical generalization"[13] [**4**, p. 974], but guided by theory. The history of science is full of examples of discoveries guided by theories, some personal and idiosyncratic, others widely held, which modern

[11] The context is a table of planets: "Each row of the table... represents an individual observation.... Each observation consists of the value of a nominal variable, which identifies the planet, and of two numeric variables, its distance from the sun, D, and the period of its orbit, P. *All* of these variables are observables" (my italics) [**36**, p. 13].

[12] An easily accessible account of Kepler's cosmology and work on the Third Law may be found in [**6**, pp. 132–147].

[13] Compare [**35**, p. 253], where Simon repeats this point, saying that the discovery of Kepler's laws "provides one of the most important and striking examples in the history of science of data-driven discovery.... *No theory* was available to explain this regularity until Newton" (my italics).

scientists no longer believe. This does not make the discoveries retroactively empirical.

Given a set of data, the BACON program does, as its creators say, use a set of fruitful heuristics and thus embody "discovery principles with a wide range of application" [**23**, p. 126]. Such a program could well outperform people in finding regularities in masses of data, allowing for experimental error, and postulating new terms, like specific heat, when needed to find such regularities. These are real accomplishments. But the major scientific discoveries BACON's authors discuss have not been made the way the program makes them.

It is instructive to contrast the reactions to Weizenbaum's ELIZA program and the BACON program on the part of their inventors. In each case, a program seems to some observers to be performing at a high level, though its accomplishments are really more modest. But Weizenbaum is appalled that people conversing with his program came to treat it like a person, while Simon and his collaborators argue that human scientists like Black, Snell, and Kepler did science the way their program does. These contrasting reactions exemplify not only two different attitudes toward the man-machine equation, but also toward the scope of individual scientific achievements.

Over and over again in the history of science one finds people involved in a major scientific breakthrough who claim that their new way of looking at things has finally settled some major moral or philosophical questions about the nature of man. In a striking parallel to the present case, Darwinian theory, in establishing its legitimacy as a mature field of science, had to contend not only with attacks on evolution from those horrified by the idea of man as an animal, but also with some Darwinists who used evolution as a basis for doctrines of white supremacy, robber-baron economics, and atheistic materialism.[14] Similarly, while some AI researchers have painted their critics as know-nothings, undervaluing or opposing all AI research, their critics have seen AI researchers as ideologues extrapolating successful researches in limited areas into a dehumanizing world-view for which there is no scientific support. Thus, Alan Turing's work—notably the abstract concept of the Turing machine—and the increasing capabilities of computing machines in the 1950s made it seem urgent to Lucas to use logical tools to defend a nonmechanistic view of man. The early successes of AI research in the 1960s produced a round of enthusiastic predictions about computers soon being able to duplicate the problem-solving and information-handling capabilities of the brain. It was these predictions and their nonfulfillment, not just the philosophy of phenomenology, that gave Dreyfus's critique its force. Again, the success of Weizenbaum's ELIZA program produced more enthusiastic predictions, and these produced in turn Weizenbaum's impassioned plea for respect for human reason. In the 1980s, Bradshaw, Langley, and Simon extrapolated from the "success" of their BACON program to claim that programs could simulate scientific discoveries and thereby duplicate some of the highest examples of human reason. The pattern described in this paragraph is, as I have argued

[14] See, e.g., [**21** and **25**]. I have discussed the generality of this pattern, together with its applicability to the recent history of AI in [**17**].

elsewhere [**17**], often characteristic of the early stages of a subject—but not of its maturity.

Meanwhile, more has been going on of a technical nature within the field of AI. Considerable research, especially since the late 1970s, has concentrated not on general problem-solving, but on programs limited to specific fields like diagnosing infectious diseases or prospecting for minerals—programs based on incorporating large amounts of systematized human knowledge. Of course, more generally-directed research continues as well. But this new direction, in what is called "expert systems," and related research in areas such as the representation of knowledge, has seen enormous growth. Expert-systems research, to quote a recent textbook, "has concentrated on the construction of high-performance programs in specialized professional domains, a pursuit that has encouraged an emphasis on the knowledge that underlies human expertise and has simultaneously decreased the apparent significance of domain-independent problem-solving theory" [**19**, p. 3]. There is overwhelming evidence for this shift in direction. One sees it from a content analysis of *Computing Abstracts* or *Computing Reviews*; by a comparative scan of successive *Proceedings* of the Joint International Conferences on Artificial Intelligence; by the appearance of new journals; and by the appearance of textbooks and collections of readings in the field of expert systems. One can see it also in many explicit statements from a new generation of practitioners, whose goal in a recent review article has been stated thus: "to provide tools that exploit new ways to encode and use knowledge to solve problems, not to duplicate intelligent human behavior in all its aspects" [**12**, p. 266]. (cp [**15**, p. 903].)

A number of these expert programs have been strikingly successful; in fact, they can sometimes outperform human experts on their designated tasks. Some measure of their success in practical terms is the great interest in business and government in developing expert systems [**46**]. Perhaps surprisingly, the very power of such knowledge-based programs strengthens the hand of critics of AI like Weizenbaum and Searle. For instance, contrast this description of an early program to devise molecular structures in organic chemistry from mass-spectrum data and empirical formulas with the claims made for BACON by its creators: "It [DENDRAL] searches for plausible hypotheses in a small subset of the total hypothesis space according to heuristic rules learned from chemists" [**5**, p. 209]. Thus, immense practical achievements, which are modest theoretically, have at least temporarily reinforced an attitude of public modesty about the nature of machine intelligence. To cite one more of a plethora of possible examples, James Albus writes in a recent text on robotics, "The answer to the question of whether machines ever will, or even can, possess a general level of intelligence comparable to humans, is unknown and may be unknowable" [**1**, p. 299].

However, some computer scientists still describe the computers of the near future as essentially involved in all aspects of intellectual life [**14**, especially part 3]; as potentially able to read newspaper accounts about terrorism and to come up with a solution to the problem [**32**, p. 220]; or as becoming conscious life forms [**37**, p. 25]. Thus, the modesty described in the previous paragraph is by no means universal.

Interestingly, much of that expert-systems research is indebted in important ways to the earlier, more general research in AI, especially in areas like the symbolic representation of knowledge and heuristic search. Perhaps the enthusiasm of many AI researchers, their desire to show that all human thought could be simulated by computer, was psychologically necessary in order to produce the successes that occurred.[15] As we have seen in the examples of Black and Kepler, the history of science abounds with examples of lasting achievements which were produced by theories and world-views no longer held. But the ideologies which produced the discoveries are not automatically validated thereby. Presumably, as AI continues to develop, it will more closely resemble a mature science. It will improve its successes in solving problems in its own sphere of competence, while ceasing to claim that it can find the ultimate truth about the nature of human intelligence.

REFERENCES

1. J. S. Albus, *Brains, behavior, and robotics*, Byte Books, Peterborough, N. H., 1981.

2. P. Benecerraf, *God, the Devil, and Gödel*, Monist **51** (1967), 9–32.

3. C. Boyer, *The rainbow: From myth to mathematics*, Thomas Yoseloff, New York and London, 1959.

4. G. Bradshaw, P. Langley, and H. Simon, *Studying scientific discovery by computer simulation*, Science **222** (1983), 971–975.

5. B. Buchanan, G. Sutherland, and E. A. Feigenbaum, *HEURISTIC DENDRAL: a program for generating explanatory hypotheses in organic chemistry*, in [**29**, pp. 209–254].

6. I. B. Cohen, *The birth of a new physics* (Revised and Updated), W. W. Norton, New York and London, 1985.

7. _____, *Franklin and Newton: An inquiry into speculative Newtonian experimental science and Franklin's work in electricity as an example thereof*, American Philosophical Society, Philadelphia, 1956.

8. M. R. Cohen and I. E. Drabkin (eds.), *A source book in Greek science*, Harvard Press, Cambridge, Mass., 1958.

9. K. Colby, J. B. Watt, and J. P. Gilbert, *A computer method of psychotherapy: Preliminary communication*, J. of Nervous and Mental Disease **142** (1966), 148–152.

10. E. J. Dijksterhuis, *The mechanization of the world picture* (C. Dikshoorn, translator), Clarendon Press, Oxford, 1961.

11. H. Dreyfus, *What computers can't do*, The Limits of Artificial Intelligence (rev. ed.), Harper, New York, 1979.

12. R. Duda and E. Shortliffe, *Expert systems research*, Science **220** (1983), 261–268.

13. D. Faust, *The limits of scientific reasoning*, Univ. of Minnesota, Minneapolis, 1984.

14. E. Feigenbaum and P. McCorduck, *The fifth generation: Artificial intelligence and Japan's computer challenge to the world*, Addison-Wesley, Reading, Mass., 1983.

15. P. Friedland, *Introduction to the special section on architectures for knowledge-based systems*, Comm. ACM **28** (1985), 902–903.

16. J. V. Grabiner, *Artificial intelligence: Debates about its use and abuse*, Historia Math. **11** (1984), 471–480.

17. _____, *Partisans and critics of a new science: The case of artificial intelligence and some historical parallels*, 1985, W. Aspray and P. Kitcher, eds., History and Philosophy of Modern Mathematics, University of Minnesota Press, Minneapolis (to appear).

18. H. Guerlac, *Joseph Black. Dictionary of scientific biography*, vol. II, Charles Scribner's Sons, New York, 1972, pp. 173–183.

[15] Besides [**17**], for the application of this generalization to the AI community, see Sherry Turkle's remarks in [**40**, pp. 251–252].

19. F. Hayes-Roth, D. Waterman, and D. B. Lenat (eds.), *Building expert systems*, Addison-Wesley, Reading, Mass., 1983.

20. J. L. Heilbron, *Elements of early modern physics*, Univ. of California Press, Berkeley and Los Angeles, 1982.

21. R. Hofstadter, *Social Darwinism in American thought*, Beacon Press, Boston, 1955.

22. T. Kuhn, *The structure of scientific revolutions* (2nd ed.), Univ. of Chicago Press, Chicago, 1970.

23. P. Langley, G. Bradshaw, and H. Simon, BACON.5: *The discovery of conservation laws*, Proc. 7th Internat. Joint Conf. on Artificial Intelligence, 1981, pp. 121–126.

24. J. R. Lucas, *Minds, machines, and Gödel*, Philosophy **36** (1961), 112–117.

25. K. Ludmerer, *Genetics and American society: A historical appraisal*, Johns Hopkins, Baltimore, 1972.

26. W. F. Magie (ed.), *A source book in physics*, Harvard Univ. Press, Cambridge, Mass., 1963.

27. P. McCorduck, *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*, Freeman, San Francisco, 1979.

28. D. McKie and N. H. de V. Heathcote, *The discovery of specific and latent heats*, Arno Press, New York, 1975.

29. B. Meltzer and D. Michie (eds.), *Machine intelligence* 4, Elsevier, New York, 1969.

30. D. E. Roller, *The early development of the concepts of temperature and heat: The rise and decline of the caloric theory*, Harvard Case Histories in Experimental Science 3, Harvard Univ. Press, Cambridge, Mass., 1950.

31. C. Sagan, *In defense of robots*, in C. Sagan, *Broca's Brain*, Random House, New York, 1978, pp. 280–292. (Article reprinted from Natural History **84** (1975), p. 10ff.)

32. R. Schanck, *The cognitive computer: On language, learning, and artificial intelligence*, Addison-Wesley, Reading, Mass., 1984.

33. J. Searle, *Minds, brains, and programs*, Behavioral and Brain Sciences **3** (1980), 417–424.

34. _____, *Author's response*, Behavioral and Brain Sciences **3** (1980), pp. 450–457.

35. H. Simon, *Computer modeling of scientific and mathematical discovery processes*, Bull. Amer. Math. Soc. (N.S.) **11** (1984), 247–262.

36. H. Simon, P. Langley, and G. Bradshaw, *Scientific discovery as problem solving*, Synthese **47** (1981), 1–27.

37. G. Simons, *Emergence of computer life*, Abacus **2** (1984), 20–25.

38. D. J. Struik, *Snel, Willebrord. Dictionary of scientific biography*, vol. XII, Charles Scribner's Sons, New York, 1975, pp. 499–502.

39. A. Turing, *Can a machine think*?, reprinted from Mind, 1950, in J. R. Newman (ed.), *The World of Mathematics*, vol. 4, Simon and Schuster, New York, 1956, pp. 2099–2123.

40. S. Turkle, *The second self: Computers and the human spirit*, Simon and Schuster, New York, 1984.

41. J. Webb, *Metamathematics and the philosophy of mind*, Philos. Sci. **35** (1968), 156–178.

42. J. Weizenbaum, ELIZA—*A computer program for the study of natural language communication between man and machine*, Comm. ACM **9** (1966), 36–45.

43. _____, *Computer power and human reason: From judgment to calculation*, Freeman, San Francisco, 1976.

44. C. H. Whitely, *Minds, machines, and Gödel: A reply to Lucas*, Philosophy **37** (1962), 62–63.

45. Y. Wilks, *Dreyfus's disproofs*, British J. Philos. Sci. **27** (1976), 177–185.

46. P. H. Winston and K. A. Prendergast (eds.), *The AI business: Commercial uses of artificial intelligence*, MIT Press, Cambridge, Mass., 1984.

47. K. Yakal, *Expert systems: Shortcut to AI*, Compute, October 1985, pp. 37–40.

DEPARTMENT OF MATHEMATICS, PITZER COLLEGE, CLAREMONT, CALIFORNIA 91711