Artificial Intelligence: Debates about Its Use and Abuse

JUDITH V. GRABINER

Department of History, California State University, Dominguez Hills, Carson, California 90747

This paper is concerned with the question, "Is what a stored-program digital computer does thinking—in the full human sense of the term?" Several current controversies are examined, including the meaning and usefulness of the Turing test to determine "intelligence." The Lucas controversy of the early 1960s is taken up, dealing with the philosophical issues related to the man-versus-machine debate, and Dreyfus' ideas against Machine Intelligence are explored. Searle's ideas in opposition to the validity of the Turing test are described, as are various interpretations of the Chinese room thought-experiment and its relation to real "thought." Weizenbaum's opposition to the "information-processing model of man" is also developed. The paper concludes with a comparison of the 19th-century debates over Darwinian Evolution and those in this century over Artificial Intelligence.

Throughout history, scientific discoveries have given rise to controversies about the essential nature of man. The debates still in progress about Darwin's account of human evolution provide an obvious example. Now, there is another. Today, we have digital computers that solve problems which, if solved by people,
would seem to require intelligence. The field which studies the computer programs involved has been given, by John McCarthy, the Aristotelian-sounding name of artificial intelligence. As an authoritative source puts it, "Artificial Intelligence (AI) is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior—understanding language, learning, reasoning, solving problems, and so on" [Barr 1981]. But is what a stored-program digital computer does thinking—in the full human sense of the term?

How could we possibly find out whether computers think the way we do? The question "Can a machine think?" has often been raised. Descartes, for instance [Descartes 1637, 56-57] said that no machine could use language as people do, responding appropriately to whatever was said to it; nor, he added, could a machine be complex enough to behave like a human in every possible circumstance. But only in our century has the question been given widespread and thorough discussion. Alan Turing suggested a way of making the question "Can a machine think?" testable without having to examine the internal complexity of the machine, by giving criteria for the intelligent behavior of computers [Turing 1950]. If, he said, a machine could successfully imitate a human being in a wide range of possible conversations—fooling its human interlocutor into believing the machine to be human—we ought to conclude that it was, indeed, thinking. The machine's success in doing this has come to be called passing the Turing test.

Turing's proposed test has set the tone for many subsequent debates. The questions most often in dispute have been these: First, will machines eventually be able to satisfy the Turing test? This is, of course, ultimately an empirical question. Second and far more important, is passing the Turing test an adequate criterion to identify real understanding on the part of a computer? And finally, what are the philosophical and moral implications of the view that the Turing test is either satisfiable or adequate? In this paper, I shall outline four controversies of the past twenty-five years, each one initiated by a different critic of the claims of researchers optimistic about the potential of artificial-intelligence research in general and the ability of computers to pass the Turing test in particular. I will discuss, as well, the way technical developments in computer science have conditioned the content of these controversies.

Turing himself expected that machines eventually would be able to pass his imitation test. It might be objected, he said, that Kurt Gödel had proved that there were limitations to the power of formal systems. Gödel's theorems show, first, that any reasonably rich formal system is incomplete, and, second, that the consistency of such a system cannot be proved within the system. Turing in 1937 had shown that formal systems are equivalent to what machines can do. Thus computers are subject to the limitations Gödel established for formal systems. Turing's reply was that people, too, may well be subject to such limitations [Turing 1950, 2109-2110]. Thus Turing, at the beginning of the modern debate over Artificial Intelligence, had already implied this startling contention: if com-
puters think the way we do, a human being is in effect an instance of a formal system.

The British philosopher J. R. Lucas was not convinced by Turing's response to this objection, and in [Lucas 1961] tried to strengthen the objection. In so doing, he began the first of the controversies we shall consider. Lucas' tools were those of mathematical logic, but his goal was much broader: here, at the outset of attempts to build thinking machines, Lucas wanted to refute what he called "mechanism"—that is, the view that the whole mind is just the sum of the operation of its separate parts, where each part's operation depends solely on its past states or on random choices between states [Lucas 1961, 126], the view that the mind is a machine. Since, as Turing had shown, a machine is just an embodied formal system, Gödel's incompleteness theorem applies to the machine. But, Lucas continued, by standing outside the consistent, incomplete formal system, we see some unprovable, meaningful formula to be true [Lucas 1961, 113]. The machine cannot produce the formula; we see the formula as true; so a human can beat every machine [Ibid., 115]. Moreover, if the human mind were nothing but an instance of a formal system, Gödel's other theorem would say that the mind could not conclude itself to be consistent. But, Lucas argues, we do in fact assert our own consistency [Ibid., 124]. Thus no mechanical model of the mind can ever be adequate.

Lucas' attempt to show that Gödel's theorem refuted mechanism provoked impassioned responses. Some simply criticized the technical aspects of his argument, asserting that he erred in his analysis and application of Gödel's theorem [Benecerraf 1967, 18; Good 1967, 357–358; Webb 1968, 156, 163, 173]. Others argued that Lucas had too limited a view of machines [Benecerraf 1967, 13–14]. Machines, for instance, which could reprogram themselves when their environments changed would not be subject to Lucas' arguments (although their potential could not be described by the study of formal systems à la Turing either) [George 1962, 62]. Most important for our purposes, some accused Lucas of holding too exalted a view of man. If a specific machine may not be able to assert the true-though-unprovable Gödel formula, people cannot always do this either [Whitely 1962, 61; Webb 1968, 168]. Finally, rather than suggest that our self-knowledge shows we are better than machines, one can reverse the argument and say that Lucas' paper really implies (since formal systems cannot know themselves) that human self-knowledge is not possible [Benecerraf 1967, 30].

The Lucas controversy, arising as it did in the early 1960s, dealt with philosophical possibilities, before actual machines had been programmed to show much intelligent-looking behavior. But, unlike the situation in Descartes' time, the pioneering programs were actually being written. As often happens in an expanding new field, as research went on, AI workers repeatedly claimed that they were close to success—success in simulating human thought. Thus the philosophical debates 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 there were any examples of machine "thought." In 1965, philosopher Hubert Dreyfus wrote a report for RAND called "Alchemy and Artificial Intelligence," which he eventually expanded into a provocative book entitled *What Computers Can't Do* [Dreyfus 1972, 1979]. In that work, after surveying a large body of then-existing AI programs, Dreyfus argued that as a matter of simple fact, not even the most highly touted AI programs should be called "intelligent" [Dreyfus 1979, 91–148]. Moreover, Dreyfus pointed out, AI researchers had vastly oversold the future potential of their work; predictions which had been made earlier (in ten years a computer will be world chess champion, discover and prove an important new mathematical theorem, translate natural languages) had clearly not come to pass [Dreyfus 1979, 81–82, 91–92]. And to these empirical criticisms, Dreyfus joined a philosophical argument: these things had not happened because they could not. But unlike Lucas, Dreyfus took his philosophical weapons not from logic but from phenomenology. AI work, Dreyfus said, rests on the false assumption that the human mind works by operating on bits of information, and performs its operation according to formal rules [Ibid., 156]. But the mind does not do this, he said. Computers, he continued, lack such human qualities as tolerance of ambiguity and the ability to distinguish between the essential and the inessential [Ibid., 107, 112]. Dreyfus asserted that man is "not a fact or a set of facts, but a being who creates both himself and the world of facts in the process of living in the world" [Ibid., 190–191]. If man is like this, then programmed computers can never satisfy the Turing test. Trying to imitate human intelligence with computers, Dreyfus has concluded, is like trying to get to the moon by climbing higher and higher in trees [McCorduck 1979, 180].

Dreyfus' views were met with outrage by many in the AI community. Seymour Papert laid Dreyfus' views to the hostility common to people who have never programmed a computer [McCorduck 1979, 196]. More substantively, practitioners of the field argued that Dreyfus' empirical description of AI work was both flawed and outdated [Wilks 1976, 177, 181; Pylyshyn 1974, 75]. Furthermore, said Yorick Wilks, Dreyfus' extrapolations about the limited future of AI research are the same kind of argument as Descartes' assertion of the impossibility of the vacuum [Wilks 1976, 177]. Moreover, the phenomenological arguments Dreyfus used to support his position are, according to his critics, neither empirically based nor intellectually rigorous (Edward Feigenbaum called phenomenology "cotton candy" [McCorduck 1979, 197]). In fact, we are 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 [Wilks 1976, 183].

The third of the controversies I will discuss began with a direct attack on the adequacy of the Turing test itself. In 1980, philosopher John Searle argued that even passing the Turing test would not mean that the machine thinks or understands. Imagine a computer program written to simulate the understanding of Chinese. Now imagine an imitation of this program by 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 squggle." 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, and so the machine he and his program are an instance of do not understand Chinese either [Searle 1980a, 417-419]. Thus the Turing test is inadequate; and the thesis that a computer with an appropriate program is really a mind, or, equivalently, the thesis that the mind is just an appropriately programmed computer, is wrong. The Turing test, said Searle, is "unashamedly behavioristic and operationalistic," and people who accept it miss the distinction between simulation and duplication [Searle 1980a, 423].

Replies to Searle's argument included cries of "sophistry" and "religious diatribe" [Dennett 1980, 428; Hofstadter 1980, 433]. Criticism of the argument itself included the statement that the total system really does contain understanding. The man may not understand Chinese, but the whole system does; if one were to look at the human brain up close, the process would look just the way the Chinese room looks to us—we are too small to see that (and how) the system understands Chinese [Lycan 1980, 435; McCarthy 1980, 435; Wilensky 1980, 450]. Some see real danger in rejecting the behavioristic approach, in which scientific models are used to represent reality [Marshall 1980, 435-436]; they warn against the "anthropocentric chauvinism" [Pylyshyn 1980, 443] that abandoning behavioral criteria might entail. Rorty [1980, 445] sees Searle's insistence on the difference between the Chinese room simulation and real human thought as equivalent to an orthodox Catholic's argument that the Eucharist conducted by a "demythologizing Tillochian theologian" or even an Anglican does not transform the wafer into the Body of Christ. In reply to such defenses of behavioral equivalence, Eccles [1980] cites an experiment by Eric Lenneberg testing the linguistic understanding of David Premack's symbol-manipulating chimpanzees. The chimpanzees successfully manipulate the symbols, but this does not mean that they have language as Premack has claimed, because, said Lenneberg, high school students, manipulating the symbols just as Sarah the chimp did (and with fewer errors), reported that they thought they were solving puzzles, and were unable to translate a single one of their completed "sentences" correctly into English (Lenneberg 1975, cited in [Eccles 1980, 430]).

Finally, some argued that Searle underestimated machines. They suggested programs embedded in robots, capable of learning from environmental stimuli not only which squggle follows which squiggle, but what happens in the world when the system makes that reply. In this situation, the programmed machine would actually learn to understand Chinese [Dennett 1980, 429]. But Searle, in his reply to his critics, accepts this as a possibility, arguing that whether this can actually be done is an empirical question [Searle 1980b, 451]. The Chinese room thought-experiment shows, Searle concludes, that merely being an instance of a formal
program is not enough; so that just passing the Turing test does not prove that the system is thinking.

The debate initiated by Searle provided much clarification of the issues surrounding the Turing test; it also placed on record some concessions on both sides: that whether or not a computer could simulate types of human understanding is a matter for empirical test [Searle 1980b, 453], and that present-day computer programs "cannot be said to be truly understanding" [Schank 1980, 445]. But asking whether the Turing test enables us to recognize another intelligence has other implications—moral ones—which are explored in detail in the last of the controversies I will consider.

In 1966, Joseph Weizenbaum of MIT developed a computer program which could simulate, in limited circumstances, a psychotherapist of the school of Carl Rogers [Weizenbaum 1966, 36–37]; (compare [Weizenbaum 1976, 3–4]). A sample of its performance follows:

**[Patient]** Men are all alike.

*["Doctor"]*

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

. . .

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

. . .

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE [Ibid.]

The accomplishment looks impressive, but is less so when the mechanism is revealed. The program responds to specified cues in specified ways: for instance, to words like "mother" with "tell me more about your family"; to statements involving "always" or "all" with requests for an example; and to statements that do not fit the predetermined script by repeating them, or, more subtly, by fishing with "does that have anything to do with" some previous topic whose importance to the patient was signaled by having been prefaced with the word "my."

Weizenbaum thought he had written a clever program; he did not dream that people would believe that the program could think. But people did. It was suggested by some that the program had passed the Turing test (as reported in [Denning 1983, 23; Boden 1977, 96–97; McCorduck 1979, 225 n., 226 n.]). It was
also suggested that such programs could practice therapy cheaply and on a large scale [Colby et al. 1966; Sagan 1975, 10]; (both quoted in [Weizenbaum 1976, 5]). Weizenbaum was shocked by these responses because they implied such widespread acceptance of the view that human beings were basically just complicated machines. In his book *Computer Power and Human Reason*, Weizenbaum characterized the information-processing model of man as just one aspect of a common 20th-century mind-set that views human beings as means, rather than as ends, and which falsely regards human problems as largely having technical solutions [Weizenbaum 1976, 11, 13–14, 251, 275]. His book is an argument that the information-processing model of man is both empirically false and 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 [Weizenbaum 1976, 208–209]. Weizenbaum characterized the most successful AI programs of the early 1970s as lacking strong theory-based models of human intelligence, being instead collections of *ad hoc* programming tricks ([Ibid., 232]). Weizenbaum concluded his critique with a call to those involved in the computer-science profession not to promulgate a view of human beings which helps to further dehumanize them. “‘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’” [Weizenbaum 1976, 277]. And again, “‘What could it mean to speak of risk, courage, trust, endurance, and overcoming when one speaks of machines?’” ([Ibid., 280].

Weizenbaum’s book, like Dreyfus’, provoked outrage and attacks on his scientific competence, like “‘He fell out of the field ten years ago, and hasn’t done a damn thing since ELIZA [the psychiatrist program]’” (quoted in [McCorduck 1979, 318]). His views on the potential of dangerous directions in some AI research have caused him to be confused with romantic critics of science and technology like Theodore Roszak, and have provoked responses to his critique of AI which are in fact general defenses of the benevolence of technology [McCorduck 1979, 326; Bell 1979, 460–461; Dertouzos 1979, 463]; and [Lederberg 1976] (quoted in [McCorduck 1979, 321]). But these responses do not invalidate Weizenbaum’s central point: thinking of people as programmed machines will affect the decisions we make about how to treat people in our technical society.

It is at present fairly widely agreed that no computer program has come anywhere near satisfying the Turing test. One might think, then, that ongoing research should have had little bearing on the debates about AI’s philosophical implications so far. But this is not the case. Turing’s logical tools—notably the abstract concept of the Turing machine—for studying computing machines, and the increasing capabilities of digital computers, are what provoked Lucas’s attempt to use logical tools to defend a nonmechanistic view of man. The early successes of AI research produced a round of enthusiastic predictions about “duplicating the problem-solving and information-handling capabilities of the brain” ([Simon 1960], quoted in [Weizenbaum 1976, 245]); their nonfulfillment and
their naïve optimism gave Dreyfus’ critique added force. Again, the success of Weizenbaum’s ELIZA program produced more utopian predictions, many from outside the computer-science community altogether, and these in turn produced Weizenbaum’s impassioned plea for respect for human reason.

Meanwhile, the direction of AI research in the 1970s began to change, from general problem solving to more successful programs limited to specific fields like medicine or geology—programs based on incorporating large amounts of systematized human knowledge. Duda & Shortliffe [1983] describe some specialized “expert systems” that can, in their limited domains like diagnosis of bacterial infections or prospecting for certain mineral ores, perform as well as human experts. Perhaps surprisingly, the power of such knowledge-based programs, whose success is based on systematizing existing human knowledge rather than simulating general problem-solving power, strengthened the hand of critics like Weizenbaum and Searle. Immense practical achievements which are theoretically modest had, at least temporarily, reinforced an attitude of modesty about the nature of machine intelligence.

However, this modesty is by no means universal (see, e.g., [Feigenbaum & McCorduck 1983, 44–45], and the report in [Waldrop 1984, 1280]). Thus there is now another division, not between involved researchers and outside critics, but between the designers of specialized expert systems and those with more grandiose aims, between those trying to make programs “smart” by building on human knowledge as necessary and those seeking programs with general intelligence [Waldrop 1984, 1279–1280].

The 20th-century debates over artificial intelligence can, I think, be better understood from the perspective of an analogy from the history of science. In establishing its legitimacy as a mature field of science, Darwinism 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. Similarly, while AI researchers sometimes paint their critics as undervaluing or opposing all AI research, their critics see AI researchers as ideologues extrapolating successful researches in limited areas into a dehumanizing world-view for which there is no scientific support. Perhaps as AI continues to mature, it will more closely resemble a science solving problems in its own sphere of competence, rather than a philosophical position claiming the ultimate truth about the nature of human intelligence.

REFERENCES


