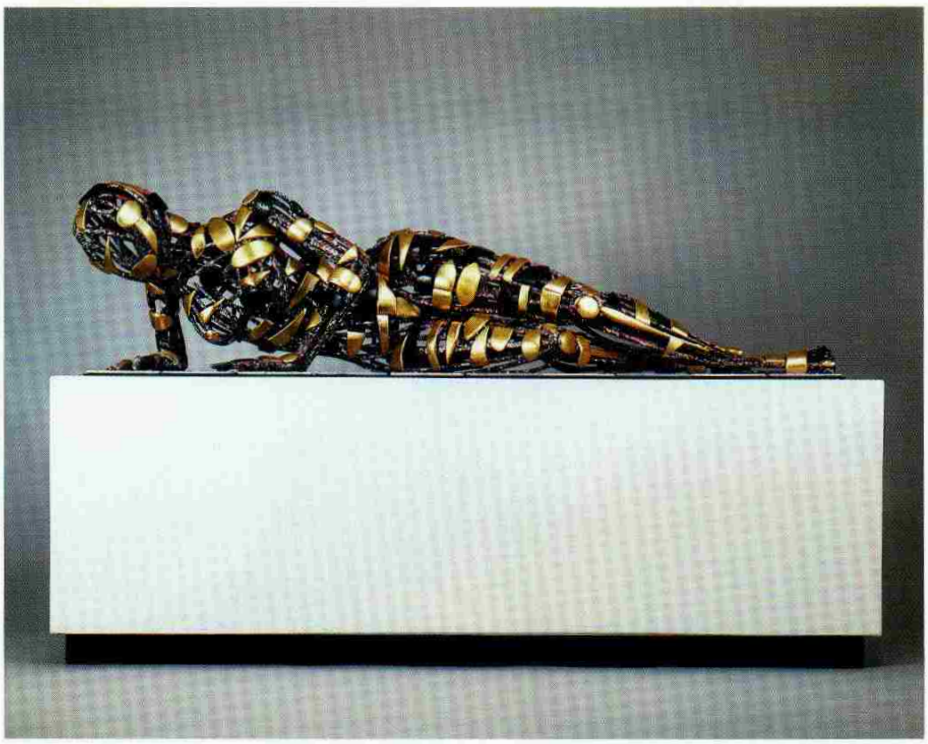
The Sciences, 27(2): 12-14, 1987.

On Human Nature

Melvin Konner



Daniel Meyer, Endangered Species IV, *1986*

Love Among the Robots

When I was a boy of sixteen or so, an episode of "The Twilight Zone" changed my life. Since I was a serious student, thoughtful and deeply religious, this rather unconventional source of a weltanschauung takes some explaining.

The show's plot was as follows: In some future society, a man has been exiled, after committing unnamed political offenses, to solitary imprisonment on an asteroid. Once a year he is visited by a supply ship, whose captain takes pity on him and brings him a present in a very large box. On opening the box after the ship has gone, the hero reacts with disgust: it is a female robot, a mere mechanical substitute for genuine companionship. The robot, however—portrayed by a most appealing actress—is remarkably lifelike and, after some fumbling and learning, appears completely human. In seeming (being?) palpably hurt by his rejections, she wins his first attention. One thing leads to another on this desolate orbiting rock, and by the supply ship's next visit, they have formed what must be construed as a powerful bond of affection. If the word has any meaning at all, they *love* each other.

The next year, the captain brings undreamed-of news; the winds of political change have blown a breath of amnesty from one end of the galaxy to the other. The hero can return home at last—but only he. The captain is smack up against his weight limit and, having failed to think of the android, has room for only one passenger. When reason fails to disabuse the hero of his sentimental attachment to the machine (the argument goes on painfully in the presence of this third party, whose face is filled with tragic apprehension), the captain resorts to his ray gun, shooting the seeming woman in the face. The wound reveals a tangle of wiring and circuitry, and the robot's voice, repeatedly calling the hero's name, runs down like a record player with its plug suddenly pulled.

As I recall the final moments of this denouement, no tears were shed, and it appeared (though this was left to the viewer's imagination) that the hero would be brought to his senses and sent home to freedom a sadder but wiser man. But to me this was murder, not only unpunished but condoned, and I could not get it out of my mind. With all the intensity of adolescent idealism, I worked the issue through in my mind and a few days later gave up my belief in the insubstantiality of the soul. By virtue of her animated responses, her full range of thought and feeling, and, above all, her trust in and love for the hero—not to mention his for her—the android was human and had as much of a "soul" as any person, regardless of whether the hardware within was carbon or silicon. Or, to put it another, more distressing way, a human being could have no more of a soul than she.

In the years since I was so shaken by this fiction it has come a long way toward fact. Artificial intelligence, then an obscure undertaking confined to a few university campuses, is now a large commercial enterprise, and, according to its enthusiastic practitioners, such as Marvin Minsky, machines of the future will not only perform bigger calculations than humans ever could but will also make medical and legal

judgments, perform psychotherapy, and compose beautiful music and poetry.

I suppose this prospect should be easier for me to accept now than it was twenty years ago. But in a way it is more unsettling than ever. For ten years or so, the relentless depredations of sociobiology—like the similarly motivated Freudian ones of an earlier era—have eroded, it seems, the very basis of the human spirit. The most cherished differences between humans and animals, one after another, have been swept aside: motherly love, altruism, cooperation, and sacrifice are now seen as mere adaptations—genetically programmed strategies for survival that we share with many other species. All that has been left to us after this beastly onslaught is rational thought: we are animals, yes, but thinking animals, and no other configuration of matter on Earth can rival us in this domain. Now even rational thought is being taken over—lock, stock, and memory board—by computers. The turf separating animal and machine is shrinking, and it is only human to wonder whether there will always be a place for us, and us alone, to stand.

The question of whether machines will ever be able to think is, in artificial intelligence—or AI—circles, commonly cast in terms of the Turing test, devised by the British computer scientist Alan M. Turing, who died in 1954. Turing imagined an "imitation game," in which a human interrogator communicates with two unseen people—a man and a woman—via teletype. The interrogator can ask any question he likes, the goal being to determine which is the man and which the woman. The catch is that the man will be trying to deceive him and is free to lie egregiously—claiming, for example, to have long, elaborately styled hair. The woman, Turing wrote, "can add such things as 'I am the woman, don't listen to him!' to her answers, but it will avail nothing as the man can make similar remarks." The questions that fascinated Turing were: What will happen when a machine replaces the man? Will the interrogator err as often as before? "These questions," he wrote, "replace our original, 'Can machines think?'"

The idiosyncrasies of this game may have had special significance to Turing, who was a homosexual, at a time when male homosexuality was a crime. (Some observers have attributed his death—an apparent suicide, involving a cyanide-laced apple—to harassment by the British government, which he had served nobly in the Second World War, breaking a critical and supposedly impregnable German code.) But the game can be recast so that it does not revolve around gender, and these days it usually is: a machine that could pass the Turing test is now defined as one that would fool a human interrogator into believing that it, too, is human.

The belief that someday a computer will pass this test—an article of faith for Minsky and like-minded computer scientists—has not gone unchallenged, of course. Humanists decry the claim that machines might think as people do, even as AI researchers try to develop machines that will decry the decrying humanists. The humanists' case rests, first of all, on what might be called the intuitional fallacy. This argument, as commonly stated, is that computers will never be able to do everything humans do, because computers rely exclusively on rules, whereas people act intuitionally, with a keen but unspecifiable sort of inference from experience. Thus, no machine will ever beat a world champion at chess, and no computer will ever be a good physician.

But here the humanists often invoke a special pleading that borders on petulance. Hubert Dreyfus, a philosopher at the University of California at Berkeley, has written about the "failures" of medical-diagnosis systems. A program known as INTERNIST-I, given laboratory test data from real case histories, missed eighteen of forty-three diagnoses, he has noted, while a team of clinicians at Massachusetts General Hospital missed a mere fifteen, and a committee of medical experts only eight. So, if you get yourself a committee of medical experts to agonize over the data, as doctors rarely would, you get slightly fewer than half as many errors as with the machine; while if you settle for a team of clinicians at one of America's best teaching hospitals, you better the program's error rate by just less than seventeen percent. It doesn't take much experience in medical practice to surmise that INTERNIST-I would probably outperform a large minority (at least) of American physicians, to say nothing of lesser-trained physicians in some parts of the world—and this at the very dawn of the use of such systems. Thus, Minsky and his colleagues properly brush aside the intuitional fallacy with allusions to the future. This aspect of what computers can't now do is technically, but not philosophically or scientifically, interesting.

The second argument humanists make, which might be called the intentional fallacy, is philosophically interesting but far from decisive. It amounts to a rejection of the Turing test. The contention is that, even if computers are someday able to accomplish the same intellectual tasks as humans, *thought* will not be the right word for the information processing behind their performance. The philosopher John Searle advanced this position in a 1980 paper, published in the journal *The Behavioral and Brain Sciences* with simultaneous replies by the great and near great of artificial intelligence and cognitive science, and with Searle's replies to the replies.

Searle's paper begins with his "Chinese room" argument: Imagine a room with no windows or doors, only a mail slot. Suppose you passed a story written in Chinese into the room, then passed in questions about the story, also in Chinese. Twenty minutes later, perfectly sensible, well-crafted answers come out of the slot, again in Chinese, suggesting that something in the room understands Chinese. But, said Searle, it might be that within the room was Searle himself, who understands no Chinese but was merely following rules he had been given for converting some kinds of foreign squiggles into other kinds. Thus, nowhere in the room is there true understanding of Chinese, even though the room behaved as if there were.

Among the criticisms of this argument that Searle did not convincingly answer are psychologist Bruce Bridgeman's—that not even humans are fully aware of the mental operations underlying conscious thought; computer scientist Douglas R. Hofstadter's—that the man in the Chinese room is functionally no more sophisticated than a few neurons, and that such a system could not possibly pull off anything so complex and subtle as language translation; and philosopher Richard Rorty's—that if the system really *could* pass this variant of the Turing test, it *would* understand Chinese, since a truly scientific definition of "understanding" must be stated in strictly behavioral terms. Further, there is the behaviorist criticism (which Searle also answers unconvincingly): our skepticism that this system has a mind implies a skepticism of other minds in general—including human ones; if we can't infer thought from behavior, we must spend our lives wondering whether anyone on the planet other than ourselves is truly conscious. This, of course, is something that few of us see a compelling reason to do.

These criticisms, taken together, suggest that the Chinese room argument is specious. But even if they did not, Searle himself has acknowledged a loophole in his argument, one that is frequently overlooked in discussions of it. He believes that mind is a kind of insubstantial secretion of the brain and emanates from our neurochemistry, as dependent on physiology as is the milk from a mother's mammary glands. A mechanical system could have mind, he concedes, if it precisely simulated the physical processes of the brain. But most AI researchers make no attempt to reconstruct the flow of information that actually occurs in a mass of human neurons—much less to build surrogates of the neurons themselves. Rather, they try to duplicate only the relation-

ship between input and output—between the slips of paper that go into the Chinese room and those that come out.

Enter the humanists' third argument, which AI enthusiasts might call the emotional fallacy, except that it isn't really a fallacy. This is the one presented by Sherry Turkle in *The Second Self: Computers and the Human Spirit*. The book is based on years of fieldwork among MIT computer hackers, AI experts and their groupies, and ordinary children playing computer games. Beginning in the late 1970s, when the age of personal computers was colorfully dawning, Turkle examined people's relationships with computers, in the twin senses of interaction and comparison. As she shows, these twins are Siamese: interaction with a computer involves an assessment, if unconscious, of how it compares to us, and comparison assumes some relationship (the Turing test, after all, implies as much).

Among Turkle's findings is that whether a machine can pass the Turing test depends on the mind of the beholder. To one five-year-old encountering it for the first time, Texas Instruments' Speak & Spell toy was alive. Other children, not convinced but clearly uneasy, took special delight in "killing" it by taking out the batteries—as if reaffirming their own uniqueness. Much more at ease with the idea of an animate machine were the hackers—college-aged people, usually men, who work, live, eat, sleep, and breathe computers. Hackers articulate frankly the satisfactions of their relationships with computers: complete devotion, predictability, and control—the kinds of things a person could never provide. As hackers themselves seem to recognize, their spirits have found in the computer a sort of superperson cut from the cloth of fantasy. The computer has passed the Turing test as posed by some of their most fundamental human needs.

But hackers are the exceptions. Most people feel the need to defend themselves from the computer's insult to their humanity. They do so, usually, by defining themselves in opposition to it: sure, the machine can play a dazzling game of chess, but only humans enjoy winning; it can diagnose illness, but only humans fear making a fatal mistake; it may have thought, but only humans have feelings. As Turkle realized, this is but a variation on the game of defining humans in opposition to animals. "Where we once were rational animals," she wrote, "now we are feeling computers, emotional machines."

We have come full circle, and our identity crisis remains unresolved. We say we are rational animals, but computers are more superbly rational; we say we are feeling machines, but other animals have the same vivid array of motives and feelings. The process of definition-by-exclusion would seem to have left us empty.

Of course, we are not. It is the intersection of the sets that makes us human—the tiny corner of the Venn diagram where animal motives overlap with mechanical rationality. It is the inner argument between the ache of sexual desire and the thought of ultimate consequence that produces the lover's plaint; the climbing of animal fear on the latticework of symbol that makes possible the comfort of ritual; the bubbling of the consciousness of our own mortality through everyday sensual experience that gives rise to the absolutely human sense of beauty.

Consider the example given by the computer scientist Joseph Weizenbaum, in *Computer Power and Human Reason*, of what computers can't simulate: the wordless communion that a mother and father share as they stand over their sleeping child's bed. Contained in their glances is the shared love growing out of the three relationships; the subtle memories of the sex that engendered the bonds; the life histories of the man and the woman—the events of their own childhoods echoing ineffably through the sleeper, the cascade of family dramas falling for generations; and, above all, the man's and the woman's sense of their own, and their child's, mortality—the fear, the grief, the intensified love of the things of this world.

Could computers simulate—perhaps even experience—this tragic sense of life? Simulate, possibly. But to experience it they would have to participate in a fully human life cycle. They would have to be born, grow, surrender themselves to some kind of family life, confront the demands of maturity, reproduce, age, and, especially, be conscious of the prospect of their dying. Not to mention their having to experience the aches and pains, the shivers and sweats, the hormonal flux, the sludge of fatigue, the neuronal dropout, and the nine-hundred-and-some-odd other natural shocks that flesh is heir to. As Turkle put it, "A being that is not born of a mother, that does not feel the vulnerability of childhood, a being that does not know sexuality or anticipate death, this being is alien."

But how well will Turkle's comforting contention fare in the future, when computers compose plainly good poems? In considering how we would respond to such poems, recall our response to the nice abstract paintings composed a few years ago by a chimpanzee. We were curious about them, admired them, even paid a good price for them, but we knew they were not real paintings. A machine much simpler than the simplest of computers could produce abstract paintings, some of which would be pleasing to the eye. But, like the chimp ones, they would not be real. Real paintings come out of human experience, respond to human traditions, are informed by human expectations even when they violate them.

Or consider poems written by children. These are often freer, more engaging, and lovelier than any the same child will be capable of writing when grown. So why don't we admire them the way we would similar ones written by adults? Because it is precisely the grown-up-ness of its source that makes the freedom and grace of a poem so admirable. A good poem by an adult is a communication from a person who, like the rest of us, has been ambushed by life but who has miraculously escaped the loss of the gracefulness that came easily in childhood—or, perhaps, has found a way artfully to recover it. In this sense a "good" poem by a computer would be of no more interest than the tragic drama typed by the random keystrokes of the proverbial roomful of monkeys—except, of course, for its value as scientific curiosity.

So what will it be like when computers are—as they will surely be—vastly smarter than we are in many ways? We will ask them, I think, to speculate about the influences of Shakespeare on Shelley, or maybe even expect them to suggest such a study, but we will not curl up near the fire with a slim volume of verse they have written. We will go to them for most sorts of medical diagnoses, maybe even for surgery. But, at our bedside, while we are dying, we will want someone who knows that he or she will also, someday, die. Computers will be, perhaps, like the gods of ancient Greece: incredibly powerful and even capable of many human emotions—but, because of their immortality, ineligible for admission into that warm circle of sympathy reserved exclusively for humans.

And what of the murdered android I mooned over at sixteen? I doubt that any robot could simulate emotion well enough to pass my ultimate Turing test: Does this machine have a tragic sense of life? Of course, my feelings toward her would not be irrelevant; relationships help define machines as they help define people, and the question of what constitutes murder hinges partly on how the murderer violates the relationship as he himself perceives it. Nonetheless, I would steel myself and apply my ultimate test; if she failed, even I might draw my ray gun. ●

Melvin Konner, who teaches anthropology at Emory University, in Atlanta, is on leave at the Center for Advanced Study in the Behavioral Sciences, in Stanford, California. His book The Healing Artisans: A Journey of Initiation into Modern Medicine, *will be published by Viking in August.*