COULD A COMPUTER EVER BE CONSCIOUS?

Steven Pinker

Steven Pinker is Professor and Director of the Center for Cognitive Neuroscience of the Massachusetts Institute of Technology and author of The Language Instinct*. This article is adapted from his forthcoming book* How the Mind Works *(Norton, October 1994).*

In one of the first episodes of the *Twilight Zone* [Season 1, Episode 7: “The Lonely” (aired: Nov. 113, 1959)] a man named James Corry is serving a fifty-year sentence in solitary confinement on a barren asteroid. Allenby, the captain of a supply ship takes pity on him and leaves behind a crate containing “Alicia,” a robot that looks and acts like a woman. Corry, of course, soon falls deeply in love. A year later Allenby returns with the news that Corry has been pardoned and that he has come to get him and a maximum of fifteen pounds of gear. Alicia, unfortunately, weighs more than that. When Corry refuses to leave, Allenby shoots Alicia in the face, exposing a tangle of smoking wires. He tells a devastated Corry, “All you’re leaving behind is loneliness.”[1]

The horrifying climax raises two vexing questions. Could a mechanical device ever duplicate human intelligence, the ultimate test being whether it could cause a real human to fall in love with it? And if a humanlike machine could be built, would it actually be *conscious*? Would dismantling it be the snuffing out of a sentient being that we felt we had witnessed on the small screen?

Pose the first question to experts in Artificial Intelligence, and you’ll get one of two answers: lifelike robots are just around the corner, or it will never happen.[2] Don’t believe either of them. These are the kinds of “experts” who a few decades ago predicted that nuclear-powered vacuum cleaners were in our future or that man will never reach the moon.[3] Certainly computers will continue to get smarter, as the recent defeat of the world chess champion, Gary Kasparov, by IBM’s Deep Blue reminds us. Today’s computers can converse in English on restricted topics, control mechanical arms that weld and spray-paint, and duplicate human expertise in dozens of areas, from prescribing drugs to diagnosing equipment breakdowns. And artificial intelligence has jumped from the laboratory to everyday life. Most people today have had their speech recognized by telephone directory assistance systems, and many have used intelligent search engines on the World Wide Web, appliances controlled by fuzzy logic chips, or mutual fund portfolios selected by artificial neural networks.[4]

Still, today’s computers are not even close to a four-year-old human in their ability to see, talk, move, or use common sense. One reason is sheer computing power. It has been estimated that the information processing capacity of even the most powerful supercomputer is equal to the nervous system of a snail — a tiny fraction of the power available to the supercomputer inside the bloated human skull.[5] But the kinds of processing are different, too. Computers find it easy to remember a twenty-five digit number, but find it hard to summarize the gist of *Little Red Riding Hood*; humans find it hard to remember the number but easy to summarize the story. One reason for the difference is that computers have a single, reliable processor (or a small number of them) working very, very fast; the brain’s processors are slower and noisier, but there are hundreds of billions of them, each connected to thousands of others. That allows the human brain to recognize complicated patterns in an instant, whereas computers have to reason out every niggling detail one step at a time. Human brains also have the advantage of sitting inside human beings, and can soak up terabytes of information over the years as the humans interact with other humans and with the environment. And brains have the benefit of a billion-year R&D effort in which evolution equipped them with cheat sheets for figuring out how to outmaneuver objects, plants, animals, and other humans.

[1] Zicree, M. S. 1989. *The Twilight Zone Companion*. 2d ed. Hollywood: Silman-James Press.

[2] Crevier, D. 1993. *AI: The tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.

[3] Cerf, C., and Navasky, V. 1984. *The Experts Speak*. New York: Pantheon.

[4] Hendler, J. 1994. High-performance artificial intelligence. *Science*, 265, 891-892. Crevier, *op cit.*

[5] Crevier, *op cit*.

So how well will tomorrow's machines do? Technological progress is notoriously unpredictable. When it comes to replacement parts for the body, who knew that artificial hips would become commonplace and artificial hearts elusive? When it comes to the performance of duplicates of the mind, the most reasonable answer is that computers will probably do a lot better than they do now, for some kinds of thinking, and they will probably not do as well as a human being, for other kinds.

But let's return to science fiction and assume that someday we really will have Alicia-class robots. Will they be "conscious"? It all depends on what you mean by the word. Woody Allen once wrote a hypothetical course catalogue with a listing for Introductory Psychology that read, "Special consideration is given to a study of consciousness as opposed to unconsciousness, with many helpful hints on how to remain conscious."[6] We laugh because we realize that the word "consciousness" has at least two meanings.[7]

One of them is Freud's famous distinction between the conscious and unconscious mind. I ask, "A penny for your thoughts?" You reply by telling me the content of your daydreams, your plans for the day, your aches and itches, and the colors, shapes, and sounds in front of you. But you cannot tell me about the enzymes secreted by your stomach, the current settings of your heart and breathing rate, the projections on your retinas, the rules of syntax that order words as you speak, or the sequence of muscle contractions that allow you to pick up a glass. This shows that information processing in the nervous system falls into two pools. One pool can be accessed by the brain modules behind verbal reports, rational thought, and deliberate decision-making. The other pool, which includes gut responses, the brain's calculations for vision, language, and movement, and repressed desires or memories (if there are any), cannot be accessed by those modules. Sometimes information can pass from one pool to the other. When we first learn how to use a stick shift, every motion has to be thought out, but with practice the skill becomes automatic (conscious processes becomes unconscious). With intense concentration and biofeedback, we can focus on a hidden sensation like our heartbeat (unconscious processes become conscious).

Will computers ever become conscious, in this sense of access to a subset of the information in the whole system? In a way, they already are. The operating system of your computer is designed so that certain kinds of information are available to the programmer or user — opening and saving files, sending messages to the printer, displaying directories — and others are not — such as the movements of the disk drive head or the codes sent by the keyboard. That's because any information system, computer or brain, has to work in real time. A device in which every morsel of information had to be easily available at all times to every process would be perpetually lost in thought. It would have to calculate whether the price of tea in China was relevant to which foot should be put in front of the other one next. Only some kinds of information are relevant to what the system is doing at a given time, and only that information should be routed in to the system's main processors. Even robots of the future, with their thousands of processors, will need some kind of control system that limits what goes into and out of the individual processors. Otherwise the whole robot would lurch and zigzag as the processors fight for control, like Steve Martin in *All of Me* when his right side was controlled by the ghost of Lily Tomlin. So in that sense, computers, now and in the future, are built with a distinction between "conscious" and "unconscious" processing.[8]

But it's a very different sense of the word "consciousness" that people find particularly fascinating. That sense is *sentience*: pure being, subjective experience, raw feels, first-person present tense, "what it is like" to see red or feel pain or taste salt. When asked to define "consciousness" in this sense, we have no better answer than Louis Armstrong's when a reporter asked him to define jazz: "Lady, if you have to ask, you'll never know."[9]

How can we ever know whether Alicia is conscious in this sense — whether there's "anyone home" seeing the world through her camera-eyes and feeling the signals from her pressure sensors? No matter how smart she acts, no matter how responsive, no matter how vehemently she

[6] Allen, W. 1983. *Without Feathers*. New York: Ballantine.

[7] Block, N., & commentators. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227-287. Jackendoff, R. 1987. *Consciousness and the Computational Mind*. Cambridge, Mass.: MIT Press.

[8] Baars, B. 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

[9] Block, N. 1978. Troubles with functionalism. In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Minnesota Studies in the Philosophy of Science, Vol. 9. Minneapolis: University of Minnesota.

says she is conscious, an Allenby can always insist that she's just a very fancy stimulus-response machine programmed to act *as if* she were sentient. Try as hard as you like, but you will not come up with an experimental test that will refute him.

Perhaps it is some consolation to know that our befuddlement here is not just a technological puzzle but is a piece with some of the deepest problems in philosophy. If I can't know whether Alicia is sentient, how can I know whether *you* are sentient? I *think* you are, and I'm not so sure about Alicia, but maybe I'm just chauvinistic about creatures that are made out of meat rather than metal. How can I be so confident that consciousness is secreted by the brain tissue in my skull, rather than lurking in the software that my brain is running — software that Alicia's computer could run just as well?[10]

Lest you think that the answer is obvious one way or another, ponder these thought experiments. Suppose surgeons replaced one of your hundred billion neurons with a microchip. Presumably you would feel and behave exactly as before. Then they replace a second one, and a third one, and so on, until more and more of your brain becomes silicon. The chips do what the neurons did, so your behavior and memory never change. Do you even notice the difference? Does it feel like dying? Is some *other* conscious entity moving in with you? Suppose that the transporter in *Star Trek* works as follows. It scans in a blueprint of Kirk's body, destroying it in the process, and assembles an exact duplicate out of new molecules on the planet below. When Kirk is beamed down, is he taking a nap or committing suicide?

The head spins in confusion; it's hard to imagine what a satisfying answer to these questions would even look like. But they are not just brain-teasers for late-night college dorm-room bull sessions. The imponderables also drive our intuitions about right and wrong. Was Allenby guilty of destruction of property, or of murder? Does a newborn boy feel pain when he is circumcised, or is his crying just a reflex? What about a lobster boiled alive, or a worm impaled on a fishhook?

These problems won't be solved any time soon, so don't expect someone to tell you with certainty whether a computer will ever be sentient. Perhaps it is a meaningless question, and we have been deluded by misleading verbiage into taking it seriously. Perhaps some unborn genius will have a thunderbolt of insight and we will slap our foreheads and wonder why the problem took so long to be solved. But perhaps the problem never will be solved. Perhaps the human mind, a mere product of evolution of one species on this planet, is biologically incapable of understanding the solution. If so, our invention the computer would present us with the ultimate tease. Never mind whether a computer can be conscious. Our *own* consciousness, the most obvious thing there is, may be forever beyond our conceptual grasp.[11]

[10] Dennett, D. C. 1991. *Consciousness Explained*. Boston: Little, Brown.

[11] McGinn, C. 1993. *Problems in Philosophy: The limits of inquiry*. Cambridge, Mass.: Blackwell.