Philosophical Conversations

An Interactive Newsletter

Vol. 6, No. 1 **Fall, 2002**

Department of Philosophy **Illinois State University**

Philosophical Conversations is designed to stimulate thought and discussion, and keep you philosophically active. The format will be the presentation of a brief position paper to which responses are encouraged. In the subsequent issues selected responses may be published in addition to a new position paper. We invite you to respond to this position paper, either by contacting the author or the Department. (Addresses, mail and e-mail, and phone numbers and fax numbers are provided at the end of this issue).

[NOTE TO THE READER: For your convenience (and further study), we have put this short essay online, on The Mind Project website, and we have included hyperlinks to some of our own webpages that will further elaborate on certain of the topics discussed here. The web version can be found at http://www.mind.ilstu.edu/events/yourobot.html]

You, Robot!?
by
David Anderson

People have long speculated about the possibility of creating machines that are as intelligent as human beings. Likewise, people have speculated about the "mechanical" nature of human beings. Are humans "nothing more than" organic machines? Could a steel and silicon machine ever acquire the properties necessary to qualify as a "person"?

Speculation about these questions is nothing new. What is new is the number of well-respected researchers in robotics, engineering, and artificial intelligence who have predicted dramatic breakthroughs in the next fifty years that will (or so we are told) answer many of these fundamental questions, once and for all. If these prognosticators are right, then human life – and "machine life" – on this planet will never be the same.

In the past decade, a number of prominent figures in areas of advanced technology have begun making two remarkable claims. The first claim is that in this century, possibly in as little as 50 years, there will be machines that are more intelligent (not to mention, more powerful) than humans. The second claim is that in that same time frame, the technology will be available to download the human "mind" and re-install it in a computer. The result is that when your biological body wears out, it will be possible to relocate your "mind" (the "real" you?) into a mechanical body with a computer brain. You will become *immortal*. Hardware can be replaced when it wears out and back-up copies of your "mind" will be available whenever the software is corrupted. And, they say, this will not happen in some distant future, but for many of you, dear readers, within *your* own lifetime.

This is heady stuff. And, as you are probably already aware, the claims being made are not merely *empirical claims* about what level of performance machines may soon achieve, but they are also robust and very controversial *philosophical claims* about what it means to be a "person," and, more particularly, what it means to be the very person that is "you." It is an empirical question whether the next two generations will produce a robot whose behavior is so complex and so subtle that it will be, quite literally, *indistinguishable* from the behavior of normal human beings. It is a philosophical question whether having the property, "behaving in such a way as to be indistinguishable from normal human beings" is *all that is necessary* (i.e., a "sufficient condition") for having the moral status of *being a person*.

Likewise, it is an empirical question whether the technology will ever be available to "read" your memories and your personality right off of the neurons in your brain so as to preserve in computational form the information contained in your memories and the behavior-patterns reflective of your personality. However, it is a philosophical question whether the resulting robot that has information about your past experiences and that shares behavior-patterns with your prior self is indeed *you* or is rather a mechanical imposter which is, infuriatingly enough, acting *as if* it is you.

Before going any further, you might wonder who the people are who are making such extraordinary claims and on what grounds they are making them. I will mention three of the most prominent people who have done much to set the course of the present debate. Hans Moravec has been a pioneer in robotics and AI

at Carnegie Mellon University for many years. He is also a favorite interview subject for many of the television programs that have charted the progress of AI and robotics during the past couple of decades. As early as 1988 he published a book, *Mind Children: The Future of Robot and Human Intelligence* where he began to defend his bold claims. In 1999 he published another book, *Robot: Mere Machine to Transcendent Mind* where he continues the story. Joining Moravec, is Ray Kurweil, an accomplished inventor and engineer who is responsible for major breakthroughs in voice recognition, synthesized speech production, to name a couple. In 1999 he received the National Medal of Technology, the nation's highest honor in technology, from President Clinton. The publication of his book, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (1999) received a good deal of fanfare in the mass media, spurred on by Kurzweil's own very impressive website (http://KurzweilAI.net) which he uses to great effect to promote and advance his views.

Moravec and Kurzweil are both well-respected in their fields and when they make bold pronouncements about the future of robotics, people listen. Yet, they are both known for being, shall we say, "enthusiastic" about their views and so their startling predictions about the future have been taken with a grain of salt by many readers. It was a matter of some note, then, when many of the same predictions were advanced by Bill Joy, co-founder and senior research scientist at Sun Microsystems. Joy is a man well-respected in the industry and considered a sober analyst of the current state of technology, not prone to flights of fancy. It came as a bombshell, then, when Joy published an article in *Wired* magazine (the cover story), titled "Why the Future Doesn't Need Us" in which he concurs with many of the claims of Moravec and Kurzweil. Joy is not so optimistic about a future filled with intelligent robots. Rather than looking forward to a time when robots will bring with them a land of plenty, he worries that they may bring our demise. Smarter and more powerful robots may care little for the well-being of humans. Joy goes so far as to suggest that computer scientists and engineers presently working on advanced AI and robotics projects are in a situation similar to the physicists working on the Manhattan project. They are faced with a moral dilemma: Should they contribute to a venture that may lead to the production of machines that will destroy the human race?

But is this really worth worrying about? Is there any reasonable chance that robots will in our lifetimes – or even in this century – become as intelligent as human beings? Is it at all plausible that you might one day *be* a robot?

Many people, including some of the authors mentioned above, too often make it seem that the main obstacle that must be overcome is the production of a computer fast enough to perform the same number of operations per second as the human brain can perform (in the firing of its neurons). But surely, creating a fast computer will be the *easiest* task to accomplish. With all the progress that has been made in AI and robotics in the past 30 years, we are still a long way from creating a computer that can perform the kind of higher cognitive functions that humans perform. There are two dominant computer paradigms -- symbol-processing digital computers and connectionist networks (see, "Two Types of Computer" at TMP). Neither is currently capable of producing anything like human-level cognitive performance. Traditional AI programs have provided us with "artificial agents" that can manipulate linguistic symbols and so, in a sense, "speak" a language (see our Iris.1 robot). But these programs can only operate "intelligently" within a very narrow range, and when you try to *scale them up* to include anything genuinely complex, their performance suffers. As Alan Turing (he of the famous "Turing Test" for machine intelligence) has argued, narrow intelligence is no intelligence at all. The very nature of intelligence is to perform rationally within a broad domain.

Connectionist networks, the second most influential computer paradigm, do better at many tasks than do symbol-processing programs. One of their strengths is pattern-recognition, which is put to good use in computer vision systems. They can be trained to make subtle distinctions between the projected images of objects, and thus (for example) can tell the difference between people who are wearing glasses and those who aren't. This paradigm too breaks down, however, when it is taken out of domains that lend themselves to pattern recognition, and when it is asked to perform the higher cognitive functions characteristic of the genuinely intelligent performance of human beings.

None of this is to say that computers will *not* achieve human-level intelligence. We are still at the very beginnings of the computer age, and it is impossible to predict what new strides will be possible with (1) variations on the symbol-processing or the connectionist network models, (2) hybrid systems that combine the strengths of both paradigms, or (3) some third computer model, like, say, one based on dynamical systems theory. But the emphasis, here, is on the "impossible to predict." I do not believe that present achievements support the view that this kind of performance can be expected at any particular time in the future.

For philosophers, though, there are many other questions that must be answered other than questions about the outward performance of computers. The claims about human immortality, and the prospects of downloading your mind into a computer, hang on far more than claims about outward performance (what we might call the machine's behavior). The issue, of course, is that a machine might perform *as if* it has beliefs, *as if* it has hopes and dreams, and *as if* it experiences pleasure and pain, and yet it may not possess any of those properties. It may be simulating them without genuinely possessing them. To determine whether a thing has a genuine mental state, we must know the fundamental nature of mental states. That requires a theory – a theory of mind, as philosophers call it. If the theory called *functionalism* is true, then machines with all of the properties that we have should be a real possibility. (For more on functionalism see our website at http://www.mind.ilstu.edu/ .. . . .). But is functionalism true? It remains a matter of considerable

controversy.

Can robots achieve human-level intelligence? Can robots possess the rich range of mental states that we do? Even if these questions are answered in the affirmative, there are further considerations that must be raised before we can have grounds for optimism about the possibility of downloading your mind into a robot and achieving immortality. If the contents of your mind . . if your memories, beliefs, and personality . . if the very person that is *you* is to be downloaded into a robot, we must have some plausible theory of "personal identity," some theory that tells us what are the necessary conditions for "being you." Just as philosophers have been arguing about theories of mind for centuries, so they have been arguing about theories of personal identity.

One of the simplest theories of personal identity is the memory theory, defended by John Locke. This might be just the ticket for those hoping for robot-immortality. For a robot in the year 2050 to be the same person that I am today, it is only required that the robot have at least one of my memories. "That seems easy," I hear you say. All that is required is that one of my memories (for example, some fact about my past that is currently stored in my brain) be encoded and stored on the harddrive of the robot's computer. But, of course, not all encoding of facts on a computer count as genuine memories. I can write the sentence "I saw John at his 6th birthday party" on the computer on which I am writing this essay. It is then a fact about my past stored on the harddrive of a computer. But it hardly follows that my laptop *remembers* seeing John at his 6th birthday party. What distinguishes, then, between the real memory of a particular person and mere stored information? That, of course, is precisely where debates about the memory theory of personal identity get interesting.

Sadly, we cannot go down that road. I have used up more than my allotted space for this discussion and so I must bring it to a close. But your reflections on these matters need not stop here. The Mind Project website has much more on this topic that will give you further food for thought. In fact, there is an entire BOOK on this subject published exclusively on The Mind Project website. Written by Winfred Phillips, while a graduate student in Applied Computer Science at Illinois State and a Mind Project researcher, Win gives a helpful overview of the authors referred to in this discussion and insights into the many philosophical issues that are raised by claims about robot intelligence and human-robot immorality. We welcome you to The Mind Project website (http://www.mind.ilstu.edu/events/yourobot.html) to read more about it.

REFERENCES:

Bill Joy, "Why the Future Doesn't Need Us" *Wired Magazine*, (April, 2000) http://www.wired.com/wired/archive/8.04/joy.html

Ray Kurzweil, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (Penguin Books, 1999). Kurzweil's website: http://www.kurzweilai.net

Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, 1988); and *Robot: Mere Machine to Transcendent Mind* (Oxford University Press, 1998).

DAVID ANDERSON joined the Philosophy Department in 1980, after earning his B.A at the University of Hawaii and his Ph.D. at Yale, and has been the bane of ISU students' existence ever since. (It's possible that he figured in the childhood nightmares of some students even before he came here.) He teaches majors courses in the history of philosophy (Western and Asian) and metaphysics, as well as general education courses in applied ethics and feminism. Most of his research has been in Buddhist and Indian philosophy, and he's presented papers on those subjects to conferences on four continents. His second book, on reductionism about persons, will be out shortly. He's particularly notorious for regularly using (or at least mentioning) the f-word in many of his classes. Emily, his cat, loved him, but then she died. Since we should probably say something positive about him, we'll add that he does continue to ride his bike to school every day. Of course being a reductionist philosopher, he's always saying that wholes like bicycles don't actually exist, only the parts are really real. So whether that little bit of ecological virtue really counts in his favor we leave to you to decide. When he isn't in his office grading papers, or at home working on his dilapidated Victorian house, or riding his bike between the two, he's usually in Paris.

Addresses

David Anderson
Department of Philosophy
Campus Box 4540
Illinois State University
Normal, IL 61790-4540
Office Phone: (309)438-7175
E-mail: dlanders@ilstu.edu
Fax: (309) 438-8028

Department
Department of Philosophy
Campus Box 4540
Illinois State University
Normal, IL 61790-4540
Department Phone: (309) 438-7665
E-mail: Philosophy@ilstu.edu
Fax: (309) 438-8028