Computer Consciousness

"I'm afraid, Dave." In Stanley Kubrick's classic film *2001: A Space Odyssey* the computer HAL uttered these provocative words as it was being shut down. They remain provocative to this day: Can a computer really experience emotions such as fear? Can it really have an *I* that is afraid to be shut down? An *I* that is the subject of genuine conscious sensations? Or must a computer, no matter how sophisticated its program and convincing its behavior, be forever devoid of conscious experiences? These are key questions about computer consciousness, questions still debated by experts and explored in blockbuster films.

The Brain As Computer

Some experts answer, "Of course a computer can be conscious. The human brain, for instance, is a computer, and it has conscious experiences. So computer consciousness is not just possible, it is commonplace."

These experts differ, however, on why, exactly, the brain can be conscious. Some are *biological naturalists*, who claim that special properties of brain biology are critical. Precisely what these properties are, and how they can generate, or be, conscious experiences, is an open question with no scientific theories yet on offer. But one implication of biological naturalism is clear: If biology is necessary, somehow, for consciousness, then any complex system that lacks biology must also lack consciousness. Since the brain is a biological computer, it can be conscious. But a nonbiological computer, like HAL, could not be conscious, no matter how compelling its utterances.

Other experts are *functionalists*, who claim that the critical properties are not fundamentally biological, but functional. The brain can be understood as running complex programs—serial, parallel, and even quantum. Certain properties of these programs are critical for consciousness. Again, no scientific theory yet explains, precisely, what these functional properties are and how they generate consciousness; perhaps concepts from information theory or complexity theory will be useful. But functionalism is clear that biology, per se, is not essential to consciousness. A nonbiological computer, like HAL, could be conscious if it is properly programmed.

Biological naturalists assert that progress in neuroscience is required to make progress in understanding consciousness. Functionalists can agree that progress in neuroscience is important, since careful study of brain function might illuminate the functional properties that are critical to consciousness. Thus both can profitably learn from neuroscience. But they debate about how this knowledge can be used. Functionalists claim that we can, in principle, use it to build conscious, nonbiological machines. Biological naturalists disagree.

Reductive and Nonreductive Functionalism

Functionalism is by far the more prevalent view among experts today. There are many versions of functionalism, and technical naunces within these versions. But functionalists can be grouped into two broad classes.

Reductive functionalists claim that mental states are identical to certain functional states: The conditions that define the different types of mental states of a system, whether

biological or not, refer only to relations between inputs to the system, outputs from the system, and other mental states of the system. The relations among inputs, outputs and mental states are typically taken to be causal relations. However, the reductive functionalist does not claim that these causal relations *cause* mental states. Instead this functionalist claims that mental states *are* certain functional states. In particular, states of consciousness are mental states and are thus, according to the reductive functionalist, identical to certain functional states. If a computer, such as HAL, happens to have the right functional states then it *ipso facto* has conscious experiences.

Nonreductive functionalists claim that mental states arise from functional organization but are not functional states. Consciousness, in particular, is determined by functional organization, but it is not identical to, or reducible to, functional organization. Nonreductive functionalism is, in one sense, a weaker claim than reductive functionalism because it claims only that functional organization determines mental states, but drops the stronger claim that mental states are identical to functional states. But in another sense nonreductive functionalism is a stronger, and puzzling, claim: Mental states, and conscious experiences in particular, are something other than functional states, and therefore have properties beyond those of functional states. This proposed dualism of properties raises the unsolved puzzle of precisely what these new properties are and how they are related to functional properties. However, the nonreductive functionalist does agree with the reductive functionalist that if a computer, like HAL, has the right functional organization then it will have conscious experiences.

Spectrum Inversion

Reductive functionalism, although controversial, is the dominant view among experts today. One thought experiment at the center of the controversy is the so-called spectrum inversion problem, which goes back at least to John Locke (1632-1704). He asked, in his 1690 *Essay Concerning Human Understanding*, if it were possible that "the idea that a violet produced in one man's mind by his eyes were the same that a marigold produced in another man's, and vice versa." Are the colors you see the same as the colors I see? More specifically, suppose that you and I are functionally identical. Would it still be possible that our color experiences differ, so that, for instance, the color I experience when viewing a ripe tomato is the color you experience when viewing fresh grass?

Functionalists, both reductive and nonreductive, answer that it is not possible for two people to be functionally identical and yet to differ in their color experiences. The reason, according to functionalism, is that every mental state, and therefore every color experience, is determined by functional organization. So if two people have the same functional organization they must have the same mental states, and therefore the same color experiences.

If it could be shown that spectrum inversion were possible, this would falsify functionalism. It would call into serious question whether computer consciousness is possible, since most arguments in favor of computer consciousness are based on functionalist assumptions. Thus the possibility of spectrum inversion is widely debated to this day.

If reductive functionalism were true, then it would in principle be possible to build a nonbiological computer, a variant of HAL, that is functionally identical to you. In

this case, if you and the computer were shown the same visual scene then the conscious color experiences of this computer, indeed all its conscious experiences, would be identical to yours.

It is likely that technology will evolve to the point where computers behave substantially like intelligent, conscious agents. The question of computer consciousness is whether such sophisticated computers really are conscious, or are just going through the motions. The answer will be illuminating not just for the nature of computers but also for human nature.

Donald D. Hoffman

University of California, Irvine

See also: Consciousness, Neuropsychology of perception, Visual perception

Further readings

Blackmore, S. (2001). *Consciousness: An introduction*. Oxford Press.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford.

Hoffman, D. (2006). The scrambling theorem: A simple proof of the logical possibility of spectrum inversion. *Consciousness and Cognition, 15,* 31 – 45.

Hofstadter, D. (2007). *I am a strange loop.* New York: Basic Books.

Pinker, S. (1999). *How the mind works*. New York: W.W. Norton.