The Problem of AI Consciousness

https://kurzweilai.net/

(credit: Susan Schneider)

Some things in life cannot be offset by a mere net gain in intelligence.

The last few years have seen the widespread recognition that sophisticated AI is under development. Bill Gates, Stephen Hawking, and others warn of the rise of “superintelligent” machines: AIs that outthink the smartest humans in every domain, including common sense reasoning and social skills. Superintelligence could destroy us, they caution. In contrast, Ray Kurzweil, a Google director of engineering, depicts a technological utopia bringing about the end of disease, poverty and resource scarcity.

Whether sophisticated AI turns out to be friend or foe, we must come to grips with the possibility that as we move further into the 21st century, the greatest intelligence on the planet may be silicon-based.

It is time to ask: could these vastly smarter beings have conscious experiences — could it feel a certain way to be them? When we experience the warm hues of a sunrise, or hear the scream of an espresso machine, there is a felt quality to our mental lives. We are conscious.

A superintelligent AI could solve problems that even the brightest humans are unable to solve, but being made of a different substrate, would it have conscious experience? Could it feel the burning of curiosity, or the pangs of grief? Let us call this “the problem of AI consciousness.”

If silicon cannot be the basis for consciousness, then superintelligent machines — machines that may outmode us or even supplant us — may exhibit superior intelligence, but they will lack inner experience. Further, just as the breathtaking android in *Ex Machina* convinced Caleb that she was in love with him, so too, a clever AI may behave as if it is conscious.

In an extreme, horrifying case, humans upload their brains, or slowly replace the parts of

their brains underlying consciousness with silicon chips, and in the end, only non-human animals remain to experience the world. This would be an unfathomable loss. Even the slightest chance that this could happen should give us reason to think carefully about AI consciousness.

The philosopher David Chalmers has posed "the hard problem of consciousness," asking: why does all this information processing need to feel a certain way to us, from the inside? The problem of AI consciousness is not just Chalmers' hard problem applied to the case of AI, though. For the hard problem of consciousness assumes that we are conscious. After all, each of us can tell from introspection that we are now conscious. It asks: why we are we conscious? Why does all our information processing feel a certain way from the inside?

In contrast, the problem of AI consciousness asks whether AI, being silicon-based, is even capable of consciousness. It does not presuppose that AI is conscious — that is the question. These are different problems, but they are both problems that science alone cannot answer.

I used to view the problem of AI consciousness as having an easy solution. Cognitive science holds that the brain is an information-processing system and that all mental functions are computations. Given this, it would seem that AIs can be conscious, for AIs have the same kind of minds as we do: computational ones. Just as a text message and a voice message can convey the same information, so too, both brains and sophisticated AIs can be conscious.

I now suspect the issue is more complex, however. It is an open question whether consciousness simply goes hand-in-hand with sophisticated computation for two reasons.

First, a superintelligent AI may bypass consciousness altogether. In humans, consciousness is correlated with novel learning tasks that require concentration, and when a thought is under the spotlight of our attention, it is processed in a slow, sequential manner. Only a very small percentage of our mental processing is conscious at any given time. A superintelligence would surpass expert-level knowledge in every domain, with rapid-fire computations ranging over vast databases that could encompass the entire internet. It may not need the very mental faculties that are associated with conscious experience in humans. Consciousness could be outmoded.

Second, consciousness may be limited to carbon substrates only. Carbon molecules form stronger, more stable chemical bonds than silicon, which allows carbon to form an extraordinary number of compounds, and unlike silicon, carbon has the capacity to more easily form double bonds. This difference has important implications in the field of astrobiology, because it is for this reason that carbon, and not silicon, is said to be well-suited for the development of life throughout the universe.

If the chemical differences between carbon and silicon impact life itself, we should not rule out the possibility that these chemical differences also impact whether silicon gives rise to consciousness, even if they do not hinder silicon's ability to process information in a superior manner.

These two considerations suggest that we should regard the problem of AI consciousness as an open question. Of course, from an ethical standpoint, it is best to assume that a

sophisticated AI may be conscious. For any mistake could wrongly influence the debate over whether they might be worthy of special ethical consideration as sentient beings. As the films *Ex Machina* and *I, Robot* illustrate, any failure to be charitable to AI may come back to haunt us, as they may treat us as we treated them.

Indeed, future AIs, should they ever wax philosophical, may pose a "problem of carbon-based consciousness" about us, asking if biological, carbon-based beings have the right substrate for experience. After all, how could AI ever be certain that we are conscious?

Susan Schneider is an Associate Professor of Philosophy and Cognitive Science at the University of Connecticut and a faculty member in the technology and ethics group at Yale's Interdisciplinary Center for Bioethics. Her work is on the nature of the self, which she examines from the vantage point of issues in philosophy of mind, artificial intelligence (AI), metaphysics, astrobiology, epistemology, and neuroscience. The topics she has written about most recently include the software approach to the mind, AI ethics, and the nature of the person. She is also a fellow with the Institute for Ethics and Emerging Technologies and the Center of Theological Inquiry in Princeton.

Her books include: Science Fiction and Philosophy: From Time Travel to Superintelligence*, and* The Blackwell Companion to Consciousness *(with Max Velmans). She is currently writing an academic book on the nature of the mind and a trade/academic book on the technological singularity.*