Artificial Intelligence, Consciousness, and Moral Status

Susan Schneider
University of Connecticut and Institute for Advanced Study

Forthcoming in *The Routledge Handbook of Neuroethics*, Syd Johnson and Karen Rommelfanger, eds.

AI and Consciousness

When philosophers ponder whether machines could be conscious, they are generally interested in a particular form of AI: AGI, or artificial general intelligence. AGI doesn't exist yet, but we now have domain specific intelligences like AlphaGo and Watson, the world Go and *Jeopardy!* champions, respectively. These systems outperform humans in specific domains, and they are impressive. But AI seems to be developing exponentially, and within the next ten or twenty years there will likely be forms of artificial general intelligence (AGI). AGI is a kind of general, flexible intelligence that can do things like make breakfast without burning the house down, while thinking of mathematics and answering the phone. Its intelligence is not limited to a single domain, like chess. Because AGIs are general, flexible, integrate knowledge across domains, and exhibit human-level intelligence or beyond, AGIs seem like better candidates for being conscious than existing systems.

Androids are already under development that are designed to look human, such as the androids intended to take care of Japan's aging population. Armed with papers about the neuroscience of human empathy, AI researchers will build robots that tug at our heartstrings. So, if and when you first encounter an AGI, will it be conscious? If you like science fiction films then you may be thinking of characters like Roy, Pris and Rachael in *Blade Runner*, or Eva in *Ex Machina* -- it seemed to *feel* like something to be them. What I'm asking is: Will it feel a certain way, from the inside, to be AGI? Will there be a subjective, *felt* quality to their mental lives? This is the problem of AI consciousness.

But why does AI consciousness even matter?

The future moral status of AI

Notice that if a being is conscious, then it seems to deserve special moral consideration, as it could suffer and feel a range of emotions. But if we've created an AGI to work for us, to force it to fight our wars or clean our homes would be akin to slavery. And destroying it would be akin to murder.

So consciousness is key to how we value AI. Conversely, it might also turn out to be key to how AI values us. For if AI is conscious, they may value us because they see in us the well of conscious experience.

From an ethical standpoint, it is imperative to know if AI is conscious, so as to avoid mistreating conscious beings. Further, a failure to be charitable to AI may come back to haunt us, as they may treat us as we treated them.

Approaching the Problem

So, how can the problem be solved? Here, I don't have a comprehensive solution, (although I have outlined an initial test (Schneider 2016). When it comes to machine consciousness, we are still taking baby steps, I suspect. One step toward any future solution is to appreciate what makes the problem so difficult.

It may initially seem that the problem is easy: Cognitive science holds that the brain is an information-processing system and that all mental functions are computations. Given this, it would seem that artificial intelligences (AIs) can be conscious, for AIs have computational minds, just as we do. Just as a text message and a voice message can convey the same information, so too, both brains and sophisticated AIs can be conscious.

However, I believe that it is an open question whether consciousness simply goes hand-in-hand with sophisticated computation for two reasons.

First, an AGI may have an architecture that bypasses consciousness altogether. For consider how consciousness works in humans. Only a very small percentage of human mental processing is conscious at any given time. 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.

Now consider that an AGI could be highly advanced, being what is called "superintelligent AI". Superintelligent AI is a hypothetical form of AI that outthinks humans in every domain – scientific reasoning, social skills, and more (Bostrom, 2015). A superintelligence would surpass expert-level knowledge in every domain, with rapid-fire computations ranging over vast databases that could occupy the space of an entire planet or encompass the entire internet. It may not need the very mental faculties that are associated with conscious experience in humans. Consciousness could be outmoded.

Indeed, superintelligent AI might rewrite its own code; a self-improving AI may opt to outmode its own consciousness or build other AGIs that are not conscious. And the processing of a superintelligence will likely be beyond the understanding of unenhanced humans in any case. It will be difficult for a human to grasp whether a given AGI system is even conscious.

Second, for all we know, consciousness may be limited to carbon substrates. 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 could also impact whether silicon gives rise to consciousness, even if they do not hinder silicon's ability to process information in a superior manner. Similar issues may arise if microchips are developed from substrates other than silicon.

In essence, we can only determine whether a given substrate is capable of conscious processing after detailed investigation of both the substrate and the larger architecture of the AI in question. Since there may be multiple kinds of AGIs (multiple intelligences, if you will), we may need to test each type of system and new substrate independently – some AGIs may be conscious, others may not be. It may be that androids that are designed to tug at the heartstrings, like Eva in *Ex Machina*, are not conscious, while some bland, unsexy server farm is.

These two considerations suggest that we should regard the problem of AI consciousness as an open question. Indeed, perhaps AGI will be pondering the same issues – *about us*. Should they ever wax philosophical, maybe they will ask if biological, carbon-based beings have the right substrate for experience. After all, how could they ever be certain that *we* are conscious?

Bibliography

Bostrom, N. (2015). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford Univ. Press.

Garland, A.(dir.) 2015. *Ex Machina*. Universal Pictures.

Schneider, S. 2016. It may not feel like anything to be an alien. http://cosmos.nautil.us/feature/72/it-may-not-feel-like-anything-to-be-an-alien [Accessed 28 December 2016]

Scott, R. (dir.) 1982. *Blade Runner*. Warner Brothers.