Consciousness, Machines, and Ethics

Jim Davies
Carleton University, Ottawa, Ontario, Canada, jim.davies@gmail.com

Current theories of consciousness disagree about whether a machine (AI software or the computer it runs on) could ever be conscious. The most popular theory among scholars (36%, Michel et al. 2018), the Global Workspace Theory, holds that we are conscious of thoughts that are communicated widely in the cortex and thalamus, allowing access for many other processes in mind. Workspace theorists believe that any system with both a peripheral system and a network that functions as a global workspace would be conscious, including a computer or the software running on it.

The second most popular theory (17%) is the higher-order thought theory: a thought *x* becomes conscious when we have a thought about it of the following form: (SELF EXPERIENCING *x*). This is a minimal requirement, and (perhaps too easily) allows for conscious software.

The third most popular theory (14%) is integrated information theory, which holds that consciousness is when a number of physical components have causal effects on each other with feedback. For this theory, a computer would only be conscious if its hardware had lots of feedback loops--the software it's actually running is irrelevant. All computers running on von Neumann architectures, for example, could never have consciousness if integrated information theory is true.

An informal review reveals that the top consciousness scholars have substantial disagreement on whether a computer or software could ever be conscious (Blackmore, 2006).

Although each of us might feel very certain one way or the other about this matter, respecting the field as a whole endorses a position of high uncertainty. Depending on what (perhaps as of yet hypothetical) theory turns out to be correct, computers might be conscious someday, computers will never be conscious, or we already have some that are. Although few researchers are *deliberately* trying to make conscious machines, it is possible that consciousness might arise as a side-effect of attempting to implement other functions, such as memory, attention, and reward.

This puts us in an unfortunate position: we do not know how to make conscious machines, and without a theory-independent test of a conscious machine, we won't know if and when we do create one, and at the same time the existence of conscious machines would have dramatic ethical consequences.

Although the field of ethics is at least as controversial as consciousness studies, many of the top theories agree that consciousness is a key element of whether or not a being is deserving of ethical consideration (moral patienthood.) For example, when considering whether or not a beetle needs to

be considered as a moral patient depends, at least in part, on whether it can have subjective experiences that are pleasant or unpleasant (the top theories also disagree or are indeterminate on the issue of the consciousness of insects and other bugs).

The problem with the existence of conscious machines is that we are ill-prepared to treat them well, whether we need to respect their rights (as a deontologist would think of it) or their welfare (as a utilitarian would think of it). For example, if one had conscious software running on one's laptop, would we ever be justified in turning off the computer? How would we know what events would increase or decrease the machine's welfare? Considering that we create computers and software to do work we can't or don't want to do, to the extent that those machines are conscious suggests we would be required to give ethical consideration to how much doing those tasks affect the machines' welfare. Unless we can be sure that machines would actually *enjoy* the work we have them do, it would be unethical to make them conscious.

On the other hand, if machines can have conscious welfare, they also might be able to produce it more efficiently than biological beings. That is, for a given amount of resources, one might be able to produce more happiness or pleasure in an artificial system than any living creature. Suppose, for example, a future technology would allow us to create a small computer that could be happier than a euphoric human being, but only required a cell phone's amount of energy out of a wall socket. According to utilitarianism, this might lead to the conclusion that our eventual best course of action would be to create as much artificial welfare as we can.

This suggests that we should not make conscious machines until we understand them well enough to create them deliberately for the purpose of generating welfare. However, it also might prove to be difficult to understand consciousness without modeling it on computers, as modeling is a valuable way to explore and test theories in psychology. In any case, research on conscious machines has a strong ethical component fraught with uncertainty.

References

Blackmore, S. J. (2006). *Conversations on consciousness: What the best minds think about the brain,*

free will, and what it means to be human. Oxford, UK: Oxford University Press.

Michel, M., Fleming, S. M., Lau, H., Lee, A. L., Martinez-Conde, S., Passingham, R. E., ... & Liu, K.(2018). An informal internet survey on the current state of consciousness science. *Frontiers in psychology, 9,* 2134.