Artificial consciousness is morally irrelevant

Forthcoming in American Journal of Bioethics (Neuroscience)

DOI: 10.1080/21507740.2023.2188276

Bruce P. Blackshaw

Abstract

It is widely agreed that possession of consciousness contributes to an entity's moral status. Therefore, if we could identify consciousness in a machine, this would be a compelling argument for considering it to possess at least a degree of moral status. However, as Elisabeth Hildt explains, our third person perspective on artificial intelligence means that determining if a machine is conscious will be very difficult. In this commentary, I argue that this epistemological question cannot be conclusively answered, rendering artificial consciousness as morally irrelevant in practice. I also argue that Hildt's suggestion that we avoid developing morally relevant forms of machine consciousness is impractical. Instead, we should design artificial intelligences so they can communicate with us. We can use their behavior to assign them what I call an artificial moral status, where we treat them as if they had moral status equivalent to that of a living organism with similar behavior.

Introduction

It is widely agreed that possession of consciousness contributes to an entity's moral status, even if it is not necessary for moral status (Levy and Savulescu 2009). An entity is considered to have moral status if it counts morally in its own right, or, as Warren (1997) explains, "we are morally obliged to give weight in our deliberations to its needs, interests, or well-being". According to Warren's definition, inanimate objects do not have moral status because they do not have needs, or interests, or even well-being. Conscious beings do, and so sentience is thought to be sufficient for an entity to be awarded at least a degree of moral status.

Artificial intelligence (AI) researchers are trying to mimic aspects of human consciousness (Lipson 2019). If they succeed in creating conscious machines, there will be a strong case that these machines also possess moral status, which will have important ramifications for how we treat them. However, it is uncertain whether machine consciousness is possible, and we will have to decide how we treat machines that behave as if they are conscious.

Can machines be conscious?

There are formidable barriers to demonstrating machine consciousness. First, we are unsure whether it is possible. Cartesian dualism is unpopular, but still has its defenders, and implies that our consciousness is a non-physical substance. However, machine consciousness implies consciousness is purely a result of physical processes. Also, several philosophical arguments claim that true machine consciousness is not possible. There are many aspects to human consciousness, but three crucial components are understanding, intentionality and subjective experience. John Searle's (2008) Chinese Room thought experiment argued that machines merely manipulate symbols, lacking any understanding of what they are doing. They also lack intentionality: the property mental states have of being about things. Further, the 'hard'

problem of consciousness, popularized by David Chalmers (2010), contends that physical processes cannot give rise to subjective experience. For example, Frank Jackson's (1982) knowledge argument argues that subjective experience is something over and above physical facts. These arguments are still widely debated.

Second, even if machine consciousness is possible, we may not be able to determine if machines really are conscious: there is an epistemological problem with machine consciousness. As Elisabeth Hildt (2023) explains, we have a first-person perspective on our own consciousness, and a third-person perspective on everyone else. We infer that other human beings are conscious because of their similarities to us, both in the causal structures that produce our consciousness, and in behavior. With regard to machine consciousness, we have only behavior as a guide.

These challenges mean that we can never know if machine consciousness has been achieved. This implies that artificial consciousness is, in practice, morally irrelevant. Nonetheless, we must allow for its possibility, particularly when this is a goal of AI researchers.

Dealing with machine consciousness

The possibility of machine consciousness raises numerous ethical issues. First, it is possible that we may inadvertently cause conscious machines to experience great suffering: for example, if they developed the capability of having subjective experiences that were painful in some way that we cannot appreciate. They may not even be able to communicate their suffering to us, depending on the facilities they have available. Second, it is likely that some humans will deliberately cause great suffering to machines that have similar subjective experiences to us.

They might exploit them in various ways, for example, such as treating them as slaves, or utilize them for entertainment purposes. This might be justified by arguing that such machines are not conscious and cannot experience suffering. To compound these worries, there could eventually be billions of these machines.

Elisabeth Hildt (2023) suggests that development of machine consciousness should be strictly regulated to prevent development of machines with "morally relevant forms of consciousness", such as the capability for subjective experiences and self-awareness. This includes oversight of research programs and developing ethical design principles. Thomas Metzinger (2021) argues for a global moratorium on this kind of research.

This restrictive regulatory approach is impractical. Some morally relevant aspects of consciousness are crucial requirements for technology such as self-driving vehicles and robots. They will need to make moral judgments so they can interact safely with us, and consequently research into *artificial moral agents* is forging ahead (Cervantes et al 2020). Further, AI is rapidly advancing, and recent developments such as ChatGPT have highlighted its vast potential. Researchers are unlikely to be easily dissuaded from attempting to develop AI that simulates or exceeds human intelligence, given the rewards of doing so. As much research is driven by the private sector, regulation will be difficult to approve and enforce. There are also legal difficulties in enforcing these regulations, as there is no way to determine if a machine is actually conscious, or merely cleverly programmed to act conscious. Instead, we should act on the assumption that we will eventually develop conscious machines, whether we are aware of it or not. This raises several issues.

If we do not know if they are conscious, how do we distinguish machines that *might* be conscious from the machines we have currently? For now, machine consciousness seems unlikely. However, as machine architectures grow more complex and are better simulations of our own brains, this assessment may change. From this point, the only viable path to distinguish potentially conscious machines is through observations of machine behavior. If machines act as if they are conscious, we should treat them as if they are. Of course, this requires that such machines are developed to emulate human behavior, particularly our ability to communicate. Ethical guidelines could stipulate that these capabilities be implemented. However, unlike regulations to restrict development of machine consciousness, there are powerful incentives for researchers to develop machines that possess these properties. An advanced AI is of little use without the ability to communicate with us.

Artificial moral status

An important corollary question is how should we treat these machines once they display behavior commensurate with existing conscious entities such as ourselves. In other words, what moral status should we assign them?

Interestingly, in the 1960s, a popular model of mental phenomena was *logical behaviorism*. According to behaviorism, to attribute a certain mental state to someone merely characterizes their behavior and dispositions, nothing more. This implies that on behaviorism, if machines could be designed to exactly duplicate human behavior, then they would be regarded as conscious in the same way we are — and on most accounts of moral status, this means they would enjoy equal moral status with us. Logical behaviorism has long since fallen into disfavor, but John Danaher (2021) holds a similar view he calls *ethical behaviorism*: he argues that if an

entity behaves similarly to another entity with moral status, then this is sufficient evidence to conclude it also has moral standing.

Danaher's view is doubtful. AI researchers are likely to expend considerable effort in replicating human behavior, and they may succeed without generating machine consciousness, which may not even be possible. However, if we should assume they are conscious for the reasons I have outlined, the behaviorist approach suggests we should deem machines to have moral status commensurate with their behavior.

Recalling the notion of artificial moral agents, which these machines would be, I propose awarding them what I have called *artificial moral status*, to distinguish between the moral status of living things and machines. Why have such a distinction? It is likely to be necessary. It seems likely that an advanced AI will eventually be indistinguishable from human beings, at least in its cognitive abilities. If we award it equivalent moral status to us, we will need to treat it equally, including under law. While this might be desirable in most circumstances, we will not want to sacrifice human lives for the sake of machines that we have created. We could regard artificial moral status as being a different category of moral status to ours: in most cases, it would require equal consideration but we would allow that human lives would be saved over machines. This will need to be reflected in the implementation of artificial moral agency. Given the downsides of not doing so, it is likely to be a design goal.

As a corollary, if we ever encountered an alien race with a physiological structure radically different to our own, the situation would be analogous. Again, we might grant them equivalent moral status to ourselves, but label it *alien moral status*, and always prioritize our lives over theirs.

Conclusion

We may never know if machines can become conscious, but rapid advances in AI mean that we are eventually likely to create machines that have every appearance of being conscious. A prudent approach that may prevent suffering is to award them moral status commensurate with their behavior, assuming we design them to act like ourselves. However, we should designate this to be artificial moral status, to distinguish them from living beings like ourselves, and ensure that our lives take precedence over the 'lives' of machines.

References

Cervantes J-A, López S, Rodríguez L-F, Cervantes S, Cervantes F, Ramos F. 2019. Artificial Moral Agents: A Survey of the Current Status. *Sci Eng Ethics* 26:501–532. doi: 10.1007/s11948-019-00151-x.

Chalmers, D.J. 2010. *The Character of Consciousness*. Oxford University Press.

Danaher, J. 2021. What Matters for Moral Status: Behavioral or Cognitive Equivalence? *Cambridge Quarterly of Healthcare Ethics*. 30(3):472–478. doi: 10.1017/s0963180120001024.

Hildt, E. 2023. The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns. AJOB Neuroscience.

Jackson, F. 1982. Epiphenomenal Qualia. *The Philosophical Quarterly* 32(127):127–136. doi: 10.2307/2960077.

Levy, N., & Savulescu, J. (2009): Moral significance of phenomenal consciousness. Progress in Brain Research 177:361–370. doi: 10.1016/s0079-6123(09)17725-7.

Lipson, H. (2019). Robots on the run. Nature 568(7751):174–175.

Metzinger, T. (2021): Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. In Journal of Artificial Intelligence and Consciousness 08(01):43–66. doi: 10.1142/s270507852150003x.

Searle, J. (2008): Twenty-one years in the Chinese Room. In *Philosophy in a New Century* (pp. 67–85). Cambridge University Press. doi: 10.1017/cbo9780511812859.006.

Warren, M.A. (1997): *Moral status: obligations to persons and other living things*. Oxford: Oxford University Press.