Shulman, C., and N. Bostrom. 2021. Sharing the world with digital minds. In *Rethinking moral status*, eds. S. Clarke, H. Zohny, and J. Savulescu, 306–326. Oxford: Oxford University Press.

AJOB NEUROSCIENCE
2023, VOL. 14, NO. 2, 79–81
https://doi.org/10.1080/21507740.2023.2188286

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Artificial Consciousness is Unlikely to Possess a Moral Capacity

Benjamin Gregg[a]

University of Texas at Austin

Elizabeth Hildt's (2023) notion of human-like artificial consciousness (AC) is vulnerable to several objections. First, she ties it to traits such as subjectivity and to capacities for rationality, intelligence, self-awareness, suffering, and sensation. But her notion of a human-like moral status displays no such traits. She simply asserts that such a status would follow from a human-like AC. But such capacities are not in themselves moral qualities, and Hildt cannot show how they would necessarily confer moral status. It is the quality of *being* a human, not the quality of *resembling* one, that persuades political communities to confer moral status on individuals.

Second, Hildt grants AC traits that would seem to allow it to construct its own morality and to confer a moral status on itself. Yet nothing she says indicates that humans would then be obliged to recognize that status—even if they acknowledged the AC's traits of rationality, intelligence, self-awareness, capacity for suffering, and sensation. In this way, among others, Hildt simply cannot eliminate the fundamental role of human agency in the possible moral status of AC.

Third, moral status is a social construct, not a feature of the natural universe. To be sure, humans could always reconceptualize the phenomenon of moral agency to incorporate a completely new understanding of technology (Verbeek 2014). Or they could decide to confer moral status on AC, perhaps by analogy to a state conferring legal personhood on corporations (Gregg 2021). That would render the moral status of AC a *metaphor* for human moral status. Like human persons, a corporation can bear responsibility even while freeing human members from corporate responsibilities. But only humans have the moral capacity to give themselves laws, primarily through legislatures, and even to author their own human rights (Gregg 2012). They can give corporations legal rights, yet corporations cannot give themselves rights. They cannot legislate or interpret legislation in legally authoritative ways. Whatever obligations corporations may have toward humans are not self-imposed but imposed by humans. Corporate personhood is instrumental, oriented on the most efficient means to achieve a given end. Whereas instrumental behavior has no capacity to evaluate the moral status of either the chosen means or a given end, normative behavior is always value-committed. It evaluates the normative acceptability of any given goal. Even as a legal person, a corporation does not pursue the value-rationality that can orient moral agency. A future AC would be no different. And as long as artifactual moral agency cannot be analogized to human moral autonomy (Johnson and Noorman 2013), it makes more sense to attribute moral responsibility to humans—who construct an AC with agency and consciousness—than to AC itself.

Fourth, AC consciousness is unlikely to be like human consciousness. Human consciousness includes values, interests, and motivations (such as not being harmed by AC) that AC need not necessarily share. More likely is that humans would construct AC at most as a "moral patient" out of a desire, say, to prevent AC's possible suffering. Even then, humans can account for the moral significance of AC without having to attribute moral agency to it (Kroes 2012). To be sure, if AC ever became like animals in the sense of being capable of experiencing pain or suffering, at that point communities might invest AC with legal

CONTACT Benjamin Gregg bgregg@austin.utexas.edu Department of Government, University of Texas at Austin, Austin, TX 78712-1704, USA.

rights to physical integrity. Or they might invest AC with legal rights to be free from the infliction of unnecessary or preventable suffering.

Fifth, moral agency is motivated. One abiding motivation of humankind is a drive to live, not only as an evolved species trait but also as a cultural artifact. On the one hand, that desire would not abate if human survival somehow came to require the extinction of AC. On the other hand, humankind will not commit collective suicide by allowing AC to preserve AC existence if doing so required the extinction of the human species. Thus humans would have reason to fear an AC that did not recognize the moral status that humans grant themselves, a status that would prohibit AC from harming humans. Tellingly, Hildt's call for steps "to avoid morally relevant forms of machine consciousness from coming into existence" can plausibly be addressed only to the humans who program AC and not to AC itself. For it is not part of human moral discourse or normative socialization. By means of her socialization, the individual internalizes the behavioral norms of her cultural environments. Her capacity for moral judgment is a capacity to ignore or override some of her normed predispositions. She can give herself rules that she later can decide to violate, for example, when the violation is ethically warranted as, say, rebellion against an unjust regime. Not so AI.

Finally, empathy as a concern for one's fellow humans is a moral deployment of consciousness (Gregg 2022). It is a thoroughly social phenomenon. The cognitive processing of experience, and the development of moral consciousness, rest on the complementary entanglement of participants' respective perspectives. Each participant is at once both communicative participant and communicative observer of other participants. Cultural and social communication is carried by individuals whose cognition reflects these cultural programs. Empathy needs to be part of these programs for social cooperation to be possible. AC cannot develop a moral consciousness if the term refers to a *social* consciousness. For to be a moral self is to be engaged with other selves; the self grasps itself vis-à-vis other selves (for example, identifies her experiences as distinctively hers). If AC has no such self, then it makes no sense to make AC legally subject to some of the regulations to which humans are subject, such as fines or imprisonment—unless AC could experience a sanction in the negative way intended by the humans who imposed it, and unless AC could be committed to obeying the law.

But AC is not a social being. It cannot act unethically in the sense of harming humans out of anger or other emotions, nor can it act ethically because of "positive" emotions such as empathy with humans (or with other animals or with other AI). The notion of an AC that would not unintentionally harm humans even if it perceived some benefit from doing so makes no sense, for AC is not able to respect or loath human beings or otherwise make judgments of value.

Emotions are biological; AI, which is not biological, has none. Hildt assumes that a non-biological artifact can have a moral status, as long as it displays a morally relevant form of consciousness. In one sense, AC is never unrelated to biology. It is invented and manufactured by biologically evolved creatures; it is a product of human culture. The plasticity of human neurophysiology makes culture possible. The biological quality of human consciousness is significant in multiple ways. Human body states intersect with human consciousness; neural configurations interact with the things we see and hear and feel. Different body states result from the subtle play of chemical and electrical signals that take place in our "brainbody." We experience these various body states as emotions (as well as drives, appetites, motivations, predispositions, moods, and phobias).

To have a brain is to have a bodily organ; to have a mind is to interact with self, others, and the environment, natural as well as social. The critical boundary between what we are as individual human beings, on the one hand, and our physical, social, and political environments, on the other, cannot be found in our brains. We humans are in part what we do, where we are, and the interactions we have with our environments by means of collective practices, deploying language and other tools. Consciousness is an achievement; it does not begin and end with the brain. And even if AC were equated with an artificial brain—an organ—it cannot be equated with mind, which involves a relationship among brain, body, and environment. If the advent one day of artificial life brings with it artificial consciousness, capable of acting upon itself, then we humans will have discovered that consciousness cannot be explained entirely in terms of neurons firing in the brain.

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Benjamin Gregg http://orcid.org/0000-0001-9510-6147

REFERENCES

Gregg, B. 2012. *Human rights as social construction*. New York: Cambridge University Press.

Gregg, B. 2021. Beyond due diligence: The human rights corporation. *Human Rights Review* 22 (1): 65–89. doi:10.1007/s12142-020-00605-x.

Gregg, B. 2022. *Creating human nature: The political challenges of genetic engineering*. New York: Cambridge University Press.

Hildt, E. 2023. The prospects of artificial consciousness: Ethical dimensions and concerns. *AJOB Neuroscience* 14 (2):58–71. doi:10.1080/21507740.2022.2148773.

Johnson, D., and M. Noorman. 2013. Artefactual agency and artefactual moral agency. In *The moral status of technical artefacts*, eds. P. Kroes and P.-P. Verbeek, 143–158. Dordrecht: Springer.

Kroes, P. 2012. *Technical artefacts: Creation of mind and matter*. Dordrecht: Springer.

Verbeek, P.-P. 2014. Some misunderstandings about the moral significance of technology. In *The Moral status of technical artefacts*, eds. P. Kroes and P.-P. Verbeek, 75–88. Dordrecht: Springer.

AJOB NEUROSCIENCE
2023, VOL. 14, NO. 2, 81–83
https://doi.org/10.1080/21507740.2023.2188288

OPEN PEER COMMENTARIES

An Immortal Ghost in the Machine?

Richard B. Gibson[a]

University of Texas Medical Branch

In their paper, Hildt (2023) surveys several socio-ethical and regulatory issues arising from research into, and the potential emergence of, artificial consciousness—synthetic beings with a claim to moral considerations comparable to existing, morally significant biological entities. After comparing several accounts of what consciousness comprises and how we would know if a machine achieved it, Hildt then explores what forms of machine consciousness would be morally relevant and what the ethical implications of these beings would be.

The scope of their analysis concerning the ethical implications of morally relevant forms of machine consciousness is brief but broad, touching upon questions of robot rights, whether emerging artificial consciousness requires "education," concerns about networked machines becoming lonely, and how to dispose of an artificial consciousness once it has expired (or, maybe, more accurately, died). It is an observation Hildt makes related to this last point, which this OPC explores.

During their ethical analysis, Hildt considers whether "once a machine has achieved morally relevant capabilities and is exercising them, it could be wrong to interrupt functioning. There could be a moral requirement to support the continued exercise of these capabilities" (Hildt 2023, 68). In other words, Hildt questions whether it would be ethically impermissible to stop a machine consciousness from functioning by withholding the raw materials necessary to continue its existence. The examples Hildt gives are of power or regular software updates. Crucially, Hildt stipulates that the ethical implications of such deprivation do not relate to the potential negative consequences to those persons who rely on such a machine consciousness. Instead, the potential impermissibility arises because ending a conscious machine would be intrinsically wrong.

Because Hildt spends much of their paper drawing comparisons to existing morally relevant beings (i.e., humans), when they present this question, it elicits a similar comparison—the deliberate withholding of materials needed for continued human survival, such as food, air, or water. In almost all circumstances, it is uncontroversial that causing someone's demise by withholding these goods is unethical and tantamount to, or even the same as, killing them. As such, if we agree with researchers such as David Chalmers (2016)

CONTACT Richard B. Gibson rbgibson@utmb.edu Institute for Bioethics & Health Humanities, UTMB, 3.102 Ewing Hall, Galveston, TX 77555-5302, USA.