Review of: "[Essay] Not Quite Like Us? — Can Cyborgs and Intelligent Machines Be Natural Persons as a Matter of Law?"

Deke Gould[1]

1 Augustana College (IL)

Potential competing interests: No potential competing interests to declare.

Review of Gervais's "Not Quite Like Us? Can Cyborgs and Intelligent Machines Be Natural Persons as a Matter of Law?"

I am grateful for the invitation to review Daniel Gervais's article on the legal consequences for our evolving concept of personhood given the emerging possibilities in artificial intelligence (AI) research. I appreciate the author's reminder, in tackling this increasingly important topic, that questions concerning personhood status are not merely matters of science fiction (p7–all page references are to the PDF copy obtained on 22 May 2023). Personhood status has already been extended by some governing organizations to non-humans and even ecosystems such as the Amazon rainforest (Coeckelbergh 2022 p129). More importantly, I strongly agree with Gervais's claim (p13), reminiscent of Wallach and Allen's claim that trained philosophers have a role to play in the development of artificial moral agents (2009, p74), that philosophers have a role to play in the development of social and legal norms regarding moral machines. I agree with Gervais's suggestion that the courts should consult recent work in linguistics, neuroscience, and philosophy (p19). And his proposal to develop tests of sapience along with a definitional approach that is rooted in a distinction between biologically-based thought and non-biological thought (p17) in order to draw a line between human and non-human sapient AI is well-worth consideration. In the following comments, I will raise one primary concern for Gervais's proposal, followed by two secondary concerns, and I will close with a list of minor textual observations that I hope will be helpful for improving future revisions of the essay.

The primary concern about Gervais's proposal is about a potential anthropocentric bias latent in a sapience test of the sort suggested here (p19), and the anthropocentric focus in much of the article. I will expand upon this concern in three ways, beginning with considerations related to Bostrom's "orthogonality thesis" (2014 p130), following with concerns about human and non-human animal comparisons, and closing with concerns from a social-relational framework. To be clear, it is clear that as a proposal for handling problems that might confront us in the near-term future, some such sapience test and definitional approach might prove pragmatic for legal contexts. But for long-term or purely philosophical purposes, we might want to revisit those anthropocentric elements of the proposal. Bostrom has put forward the orthogonality thesis: instrumental rationality and final goals are largely independent, and represent orthogonal axes. If this thesis is correct, then the space of possible superintelligent entities extends well beyond the niche that human minds occupy. Additionally, we might consider the many different independent dimensions of cognition to observe that, as our minds are the products of an evolutionary history full of many contingent biological factors, human minds likely occupy a small slice of the map of

possible minds. Some have argued that thinking along such lines might provide moral reason to (a) abstain from creating future minds like our own, and (b) to make minds very different from our own (e.g., Gould 2021). If the space of possible future minds extends well outside the range that our contingent evolutionary lineage confines biological minds to, and if there are reasons to not subject future minds to the sorts of suffering that human minds typically endure, then some moral considerations might provide reasons to be cautious about assuming an anthropocentric focus when determining moral status. Furthermore, if we should be cautious about assuming such an anthropocentric focus in determining moral status, then we should be wary of the claim that "there is an *ought* that separates homo sapiens and machina sapiens" (p15). One might object that such an approach might beg the question against alternatives that are more inclusive of other potential future minds.

Moreover, some of the comparisons between human and non-human animals used to advance the case throughout the article might be challenged. The term "sapience" is defined as "the unique way in which reason and emotion interact in our brain and body and that, according to Darwin (and many others after him), puts us at the pinnacle of the animal kingdom" (p5). Yet this sort of hierarchical view of the difference between human and non-human animals has been challenged by many in philosophy of mind and cognitive ethology (e.g., Allen and Bekoff 1997). On some views, human capacities for complicated social behaviors such as play or vigilance belong on a continuum with non-human species, not in a mere hierarchy. Thus, one might challenge the background that underlies the claim that we have a "capacity to create a 'social reality' that is unique in the animal kingdom" (p10). While there might be reason to adopt a view that examines features of human social behavior that are unique to human beings, it is also important to acknowledge that there are reasons to resist a hierarchical picture.

Finally, some philosophers have challenged anthropocentric moral and political theory by means of promoting alternative "social-relational" frameworks. Coeckelbergh, for example, has recently offered a "cooperative-scheme" approach for thinking about political obligations to non-humans (2022 p128). On such a view, moral and political status is afforded by an entity's cooperative role in the broader social-relational framework, and not merely as a matter of cognitive or linguistic capacities. Social-relational views such as this pose a challenge to test-based approaches. For, on a social-relational framework, political or ethical status isn't afforded merely on the basis of performance on some measure of ability, but on their participation in the broader social-relational system (Coeckelbergh 2022 p142). To be clear, some proponents of the social-relational view are skeptical of the possibility that AI systems might pass the relevant tests for sentience or sapience. Nevertheless, they insist that some entities might still be viable candidates for moral and political consideration (Coeckelbergh 2022 p128).

I will end the substantive part of these comments by raising two secondary concerns, and I will include a list of some typographical and other minor issues after the "works cited" section below. The first of the secondary concerns relates to the gray area posed by the concept of cyborgs. Gervais is clear that the primary focus is the biological function of human cognition for drawing the line between homo sapiens and machina sapiens. A human with an artificial limb is still clearly a human in the ways that count (p18). However, as Gervais acknowledges, what matters is what happens in the human

brain. Some alterations to the human brain might become relevant differences. To borrow Schneider's gradual brain replacement thought experiment (2019 p26), suppose that a volunteer had each region of their brain gradually replaced with a functional isomorph such that the volunteer's "mind map" is preserved. In such a scenario, relevant questions about both personal identity and the persistence of consciousness might emerge. That said, I would urge that the gray area emerge now, long before we find ourselves confronting these sorts of distant technological scenarios. For as Clark and Chalmers (1998) and later Clark (2003) argued, we have reason to take the extended mind thesis seriously, that is, that it is a typical, natural part of human cognition to incorporate materials from outside of our "skin-bag" (Clark 2003 p27). If such a view is correct, then the issue isn't merely a matter of whether some percentage of our cognition is performed by a biological brain and which percentage isn't (p8). Maybe the best way to understand humans and human cognition is by starting with the recognition of our natural capacity to incorporate technology from the outside of our skulls, whether that is by means of marks on physical objects or by means of some directly connected electronic components. In any case, the line between human and non-human sapience might turn out to be blurrier from the start than Gervais portrays it here.

The last secondary point concerns how Gervais's proposal compares to similar tests. Labossiere (2017) and Schneider (2019) both proposed tests of cognitive capacities reminiscent of Gervais's proposal, but might provide fruitful rival perspectives that could help sharpen the advantages or disadvantages of his approach. Labossiere argues that there are reasons to take seriously an "ersatz" approach when it comes to assigning moral status (2017 p302). Such a proposal might offer additional support for Gervais's distinction between homo sapiens and machina sapiens, or it might offer an additional opportunity for clarification by way of contrast. Schneider's "AI Consciousness Test" is different in that it suggests a way for determining whether an AI system is conscious and consequently a path forward for attributing moral status (2019 p51). Schneider's AI Consciousness Test has its critics (e.g., Udell and Schwitzgebel 2021), but it seems a worthy challenge to any proposal that aims to provide a means for drawing a line between human and non-human AI persons.

Works Cited

Allen, C. and M. Bekoff. (1997) Species of Mind. MIT Press.

Bostrom, N. (2014) Superintelligence. Oxford UP.

Clark, A. and D. Chalmers. (1998) "The Extended Mind." Analysis. 58.1: 7-19.

Clark, A. (2003) Natural Born Cyborgs. Oxford UP.

Coeckelbergh, M. (2022) The Political Philosophy of AI. Polity.

Gould, D. (2021) "Future Minds and a New Challenge to Anti-Natalism." Bioethics. 35.8: 793-800.

Labossiere, M. (2017) "Testing the Moral Status of Artificial Beings." Robot Ethics 2.0. Lin, Jenkins, & Abney (eds.). Oxford UP. 293-306.

Schneider, S. (2019) Artificial You. Princeton UP.

Udell, D. and E. Schwitzgebel. (2021) “Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed.” Journal of Consciousness Studies. 28.5-6: 121-144.

Wallach, W. and C. Allen. (2009) Moral Machines. Oxford UP.

Minor Typographical and Other Writing Issues:

1. p2: typo–”Lert us also explain…”
2. p3: “This is a whlly different question.”
3. p3: incorrect use of the logical phrase “begs the question”
4. p3: missing space in “anatural person”
5. p4: “Though sentience affects sapience, the latter…” did the author mean “former”?
6. p5: missing space in “theAmerican Heritage”
7. p6: missing space in “personbehaves”
8. p6: “Human that person remains”
9. p6: “acobot” missing space
10. p7: “thetemporal” missing space
11. p8: “the mereability” missing space
12. p9: “like usand then some” missing space
13. p10: “importance ofconcepts” missing space
14. p10: “able to constructgoal-based” missing space
15. p11: “biological body plays” the letters a and y are merged
16. p12: “VII. ELements” typo in the second letter of “elements”
17. p13: “(Russell, 2019, 26)” I couldn’t find this source in the works cited section
18. p13: “Works in philosophy of mind suggests that both humans and machines thinking are ways…” grammar
19. p14: “but no makes another interesting” grammar
20. p15: “withreasons” missing space
21. p15: “cannotbe” missing space
22. p16: “VIII OTher” typo
23. p17: “featurefrom other”
24. p17: “betweenhomo spapiens”
25. p18: “but ofenhanced”