From AI for People to AI for the World and the Universe

Seth D. Baum and Andrea Owe
Global Catastrophic Risk Institute
https://gcri.org

AI & Society, 38(2): 679-680, April 2023, DOI 10.1007/s00146-022-01402-5.
This version 10 July 2023.

Abstract
Recent work in AI ethics often calls for AI to advance human values and interests. The concept of "AI for people" is one notable example. Though commendable in some respects, this work falls short by excluding the moral significance of nonhumans. This paper calls for a shift in AI ethics to more inclusive paradigms such as "AI for the world" and "AI for the universe". The paper outlines the case for more inclusive paradigms and presents implications for moral philosophy and computer science work on AI ethics.

Keywords: AI ethics; AI for people; environmental ethics; nonhumans

New Paradigms for AI Ethics

In recent years, a variety of AI ethics projects have developed frameworks intended to ensure that AI advances human values and interests, including "human-compatible AI" (Russell 2019), "AI for humanity" (Braunschweig and Ghallab 2021), and "AI for people" (Floridi et al. 2018). These projects make important contributions both by highlighting the ethical significance of AI technology and by emphasizing that AI should serve the public interest instead of the narrow private interests of developers. However, these projects fall short in that they formulate AI ethics exclusively in human terms. AI should benefit more than just humanity. It should benefit the entire world, and, to the extent possible, the entire universe.

Humans are of course a special species, but humans are not supernatural. Humans are part of nature, members of the animal kingdom and of ecosystems. Humans are not the only species with morally relevant attributes such as the capacity to form preferences, experience pleasure and pain, or have a life worth living. Indeed, in consideration of these and other factors, some have argued for the intrinsic moral value of all living beings (Taylor 1986), ecosystems (Rolston 1988), and even of abiotic environments (Milligan 2015). Similarly, it may also be the case that artifacts humans create, including AI, could have intrinsic moral value. This is especially likely for more advanced future forms of AI, such as brain emulations (Sandberg 2014). These various specific positions are all open to scientific and moral debate. What should not be up for debate is that humans are not the only entities of intrinsic moral significance, meaning that they should be valued for their own sake and not just in terms of their value for humans.

We therefore call for AI ethics to include nonhumans—not just as tools that can be used to benefit humans, but as entities that themselves have intrinsic moral value. Some work in AI ethics already does this, especially the active line of research on the rights of robots and AI systems (Gunkel 2018). But too much of the field does not. Recent analysis finds nonhumans were addressed in only 8 of 84 sets of AI ethics principles and only 8 of 34 chapters of AI ethics collections (Owe and Baum 2021a), and that nonhumans are treated as having intrinsic moral significance in only 10 of 82 publications on AI and

sustainability (Owe and Baum 2021b). This is better than nothing, but it is not enough. Nonhumans should receive consistent and robust attention in AI ethics.

Much is at stake. AI technology can have large effects on nonhumans. AI technology already requires significant material and energy inputs, especially for training large neural networks (García-Martín et al. 2019), which has harmful effects on nonhumans; this may only increase as AI systems use more computing power. AI also can be used to address issues such as global warming (Rolnick et al. 2019), which could have positive effects on nonhumans. An underexplored possibility is that machine learning systems could perpetuate biases against nonhumans in the way they perpetuate certain biases within human populations (Owe and Baum 2021a). Finally, if future AI technology ever reaches the extreme point of leading to runaway superintelligence, with humanity no longer in control (Bostrom 2014), then the value of the outcome could be highly sensitive to whether the AI is designed to include the intrinsic value of nonhumans. In sum, a large and potentially astronomical amount of moral value could be lost if AI ethics continues to neglect nonhumans.

A twofold effort is needed. First, philosophical work in AI ethics should explicitly acknowledge the moral significance of nonhumans and incorporate it into ongoing work. For example, sets of AI ethics principles should include statements on nonhumans. One proposed statement is, "The main objective of development and use of AIS [AI systems] must be to enhance the wellbeing and flourishing of all sentient life and the natural environment, now and in the future" (Owe and Baum 2021a). Other activities of AI ethics, including research studies, and advising AI projects, should also incorporate the moral significance of nonhumans. Finally, the names of AI ethics initiatives should reflect the intrinsic value of nonhumans. Names like "AI for people" are good; names like "AI for the world" or "AI for the universe" are better.

Second, computer science work in AI ethics should study and develop algorithmic techniques that account for the intrinsic value of nonhumans. Techniques that rely on observing human behavior to infer human values (Russell 2019) may not readily apply to learning nonhuman values. Research on alternative techniques would also be of value for humans, considering both the limitations of behavioral observations as a source of insight into values in general (Sen 1973) and the impossibility of observing the behavior of humans who do not yet exist. Proxy schemes have been proposed for incorporating the values of nonhumans and future humans into AI systems (Baum 2020). Further work is needed to assess how proxy or other schemes could be implemented in actual AI systems.

This is an exciting time in AI ethics. The field is awakening to a richness of possibility and exploring a great range of new ideas. It is entirely understandable that some early work has focused on humans and neglected nonhumans. We are optimistic that this can prove to be a temporary oversight, to be corrected as the field of AI ethics acknowledges that both humans and nonhumans have intrinsic moral value.

Acknowledgments
We thank Robert de Neufville and an anonymous reviewer for helpful comments on an earlier version of this paper. Any remaining errors are the authors' alone.

References

Baum SD (2020) Social choice ethics in artificial intelligence. AI & Society 35(1):165–176
Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford
Braunschweig B, Ghallab M (Eds.) (2021) Reflections on Artificial Intelligence for Humanity. Springer, Cham, Switzerland
Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B (2018) AI4People—an ethical framework for a good AI society:

opportunities, risks, principles, and recommendations. Minds and Machines 28(4):689–707
García-Martín E, Rodrigues CF, Riley G, Grahn H (2019) Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing 134:75–88
Gunkel DJ (2018) Robot Rights. MIT Press, Cambridge, MA
Milligan T (2015) Nobody Owns the Moon: The ethics of space exploitation. McFarland and Company, Jefferson, NC
Owe A, Baum SD (2021a) Moral consideration of nonhumans in the ethics of artificial intelligence. AI & Ethics, 1(4):517–528
Owe A, Baum SD (2021b) The ethics of sustainability for artificial intelligence. Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI (CAIP 2021): 1-17
Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, Ross AS, Milojevic-Dupont N, Jaques N, Waldman-Brown A, Luccioni A (2019) Tackling climate change with machine learning. https://arxiv.org/abs/1906.05433
Rolston H III (1988) Environmental Ethics: Duties to and values in the natural world. Temple University Press, Philadelphia
Russell S (2019) Human compatible: Artificial intelligence and the problem of control. Penguin, New York
Sandberg A (2014) Ethics of brain emulations, Journal of Experimental & Theoretical Artificial Intelligence 26(3):439–457
Sen A (1973) Behavior and the concept of preference. Economica 40(159):241–259
Taylor P (1986) Respect for nature: a theory of environmental ethics. Princeton University Press, Princeton