PSYCHOLOGICAL INQUIRY
2023, VOL. 34, NO. 2, 100–101
https://doi.org/10.1080/1047840X.2023.2248857

Routledge
Taylor & Francis Group

COMMENTARIES

Research on Artificial Intelligence is Reshaping Our Definition of Morality

Zoe A. Purcell[a] and Jean-François Bonnefon[b]

[a]Artificial and Natural Intelligence Toulouse Institute, University of Toulouse, Toulouse, France; [b]Toulouse School of Economics and Quantitative Social Sciences, Centre National de la Recherche Scientifique (TSM-R), University of Toulouse, Toulouse, France

Some claims about morality cannot be tested without defining morality in the first place, that is, without providing a clear and consensual set of criteria allowing to decide whether a process, decision, judgment or emotion qualify as moral or not. For example, if we seek to argue that morality is universal by showing that some process, X, is universal, we need X to clearly and consensually qualify as a moral process. If we seek to show that some moral emotions do not stem from intentional harm, we need to have a clear and consensual way to decide which emotions qualify as moral. There are comparable questions in the corner of moral psychology that intersects with Artificial Intelligence (AI), questions that focus on a contrast between the moral and the non-moral, and accordingly require a consensual definition of what separates them. For example, we may want to show that people are especially averse to AI making moral decisions (Bigman & Gray, 2018), more than they are averse to AI making non-moral decisions. To test this claim, we need a definition of what separates moral from non-moral decisions, as a required step for building appropriate experimental materials. The alignment problem may also fall in this category: if we want to align the behavior of AI to the moral values and priorities of humanity, we arguably need to agree about the boundaries of these values and priorities, in order not to complicate the alignment process by including non-moral values and priorities (Mittelstadt, 2019).

In contrast, some research questions that stem from the development of AI inform rather than require a definition of morality. Consider the following question: to what extent and for which reason do people demonstrate prosocial behavior toward AI-powered machines (Pauketat & Anthis, 2022)? When it is directed toward other humans, prosociality typically involves actions and emotions that spring from concerns about the welfare of others, which makes it a moral issue. Is prosociality toward machines a moral issue? According to the definition offered in the target article (Dahl, this issue), it is not, because morality must involve concerns about *sentient* others. As a result of this definition, we would consider that our theories of moral psychology do not have to (and maybe should not even try to) explain the prosocial behavior that people display toward machines. The problem is that we do not know yet whether this is a reasonable conclusion, because we do not know yet whether we need to extend our definition of morality so that it includes concerns toward AI as a special class of non-sentient moral patient. More generally, the development of AI blurs the boundaries of moral psychology by potentially introducing a novel class of moral agents and patients (Bonnefon et al., In Press). We do not fully know if and in which sense humans think of AI-powered machines as moral agents whose behavior they should praise or punish, or if humans think of AI-powered machines as moral patients whose welfare they are obligated to attend—but we should not exclude these questions from the purview of moral psychology solely because AI-powered machines are not sentient. We should rather embrace the fact that these questions aim at redefining what counts as morality, and respecifying what theories of moral psychology must be able to explain.

Perhaps more controversially, we need to consider that moral psychology may have to increase its scope to study machine participants. This may seem like a far-fetched idea. After all, if moral psychology is concerned with explaining human thinking and behavior, then moral psychologists learn nothing of value by running surveys and experiments with machine participants—but the goal of these experiments is not to better understand the human mind. Running experiments on machines is a key element of Machine Behavior (Rahwan et al., 2019), an emerging interdisciplinary scientific field that focuses on studying the actions and decision-making processes of AI systems. It seeks to understand how these systems interact with their environments, other machines, and human beings, as well as the underlying principles that govern their behavior—in particular, the moral principles that govern their moral behavior. Importantly, the moral behavior of machines and the principles that govern it are often difficult, if not impossible, to predict from their code alone, especially in the case of deep learning systems. First, the behavior of these systems is affected by the interaction of millions of parameters learned during their training phase, leading to emerging properties which were never explicitly encoded. Second, these systems are deployed in dynamic settings, when they encounter situations and interaction partners which were not part of their training data, making it even more difficult to anticipate their behavior in the wild. As a result, researchers who seek to understand what moral decisions the system will make, and why, need to conduct experiments in which the

CONTACT Jean-François Bonnefon jfbonnefon@gmail.com Toulouse School of Economics and Quantitative Social Sciences, Centre National de la Recherche Scientifique (TSM-R), University of Toulouse, Toulouse, France.

system acts as a participant, just as behaviorists conduct experiments with humans and other animals.

Conducting experiments to investigate the moral principles and behavior of machines may not currently fall within the purview of moral psychologists, under any definition, as their field primarily focuses on human thought and behavior. However, moral psychologists possess unique qualifications that make them well-suited for this task. They are experts in the behaviors that machines are trained on, have a deep understanding of human biases that may have influenced the machines during development, and possess the methodological skills necessary to design experiments that explore the mechanisms underlying moral behavior. Accordingly, excluding moral psychologists from the study of machine behavior would be a significant loss. However, they may be discouraged from participating if their definition of morality does not encompass machine thought and behavior, as they might perceive such work as too risky for publication in the top journals of their field. Furthermore, even if some moral psychologists choose to engage in this research, excluding machine behavior from the definition of morality would risk delegitimizing their contributions in the eyes of policymakers and scientists from other disciplines, since it would be perceived as an admission that machine behavior is not in the scope of expertise of moral psychologists.

This is in fact a broader problem, which applies beyond the specific case of Machine Behavior. Because the field of AI moves very fast, new problems keep on emerging, which may benefit from the perspective of moral psychologists, without falling squarely within the purview of morality as currently defined in their field. The list could be endless, but for example: How should autonomous vehicles distribute risks across road users (Bonnefon et al., 2016)? Should machines be allowed to pass as humans if this deception has beneficial effects (Ishowo-Oloko et al., 2019)? Should we regulate algorithmic moral profiling (Purcell & Bonnefon, 2023)? When should people be allowed to use AI to transform the way the sound, look, or read to others (Hancock et al., 2020)? How is responsibility apportioned when humans and machines jointly contribute to a harmful outcome (Awad et al., 2019)? All these problems invite a multidisciplinary approach involving computer science, engineering, law, political science, ethics, and yet other fields. But do moral psychologists get to sit at the table, if these problems do not quite fit their current definition of morality? In such a case, moral psychologists may attempt to change their definition of morality so that it encompasses the new problems they want to tackle—but this will be a slow process, slower than the pace at which new problems are created by AI. They may also start working on these new problems even though they do not quite fit their current definition of morality—but this may fragilize their position in the eyes of other scientists and policymakers, since, by their own admission, they would be working outside of their primary field of expertise.

We acknowledge that these are not arguments about science, but arguments about scientific institutions and incentives. The target article makes a compelling case that being clear about our definition of morality will improve theories of moral psychology. We agree—but as we argued in this commentary, advances in AI are disrupting our conception of morality in a way that reshapes its definition, and keep creating new problems and opportunities for moral psychologists, which may require us to keep our definition of morality as fluid as possible in order not to gatekeep ourselves out of novel, critical fields of investigation.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The authors are supported by grant ANR-19-PI3A-0004, grant ANR-17-EURE-0010, and the Research Foundation TSE-Partnership.

References

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2019). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), 134–143. Article 2. https://doi.org/10.1038/s41562-019-0762-8

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonnefon, J. F., Rahwan, I., & Shariff, A. (In Press). The moral psychology of Artificial Intelligence. *Annual Review of Psychology*.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science (New York, N.Y.)*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Ishowo-Oloko, F., Bonnefon, J. F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521. Article 11. https://doi.org/10.1038/s42256-019-0113-5

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. Article 11. https://doi.org/10.1038/s42256-019-0114-4

Pauketat, J. V. T., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior*, *136*, 107372. https://doi.org/10.1016/j.chb.2022.107372

Purcell, Z. A., & Bonnefon, J. F. (2023). Humans feel too special for machines to score their morals. *PNAS Nexus*, *2*(6), pgad179. https://doi.org/10.1093/pnasnexus/pgad179

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y