Editorial to the Special Issue on Morality and AI

Jean-François Bonnefon[1*]

[1]Toulouse School of Economics, CNRS (TSM-R), University of Toulouse Capitole, Toulouse, France

*Correspondence should be addressed to: jean-francois.bonnefon@tse-fr.eu

Artificial Intelligence is a transformative technology that may well change every aspect of our lives and societies, in ways which are currently hard to anticipate (Capraro et al., 2024). But anticipate we must. Faced with the prospect of major transformations, we need to make every effort to steer the development and deployment of intelligent machinesso as to increase their positive impact and mitigate their negative impact. This requires us to ask some hard questions about ethics and morality: What moral code should we embed in machines? How do we judge whether machines comply with this code? And could life among intelligent machines, in turn, rewire *our* own moral code? Cognitive and behavioral scientists have a frontline role in answering these questions (Bonnefon et al., 2024), and the papers in this special issue tackle them all.

When it comes to defining the moral code of machines, it is often insufficient to provide lists of ethical principles that machines should pursue (such as beneficence, transparency, and respect for autonomy), because machines will need quantitative guidance on what to do when these principles come into conflict (Mittelstadt, 2019). As a result, if we want machines to make moral decisions in complicated situations, in a way that aligns with what humans want, we need to do the cognitive and behavioral work to identify what humans actually want in these situations (Awad et al., 2018). Three articles in this special issue examine the moral values people would like to be embedded in machines. Liu and colleagues focus on the ethical dilemmas of autonomous

vehicles, and examine in particular the gap between what people think is morally acceptable from a human driver in a dilemma situation, and what they think is morally acceptable from a machine driver (Liu et al., 2025). Myers and Everett investigate the reluctance that people have for taking moral advice from machines that they expect to have utilitarian values, and how they expect to disagree with these machines in the future even when they agree with a given advice (Myers and Everett, 2025). Finally, Purcell and colleagues report that demographic homogeneity in the current AI workforce means that AI 'builders' tend to favor moral values for machines that diverge from those of the general population—and that boosting workforce diversity in the AI sector would be a promising step toward realigning these preferences (Purcell et al., 2025).

Once moral machines are developed and deployed, we must judge what they actually do (Hidalgo et al., 2021),since this behavior is not always predictable (Rahwan et al., 2019)—and these judgments involve complex psychological assessments about machine agency and patiency (Ladak et al., 2024). Four articles in the special issue deal with our perceptions of machines. Reinecke and colleagues explore how children currently perceive the moral standing of robots, a critical step for us to anticipate how future generations will relate to intelligent machines, having grown up in a world where they are commonplace (Reinecke et al., 2025). Two articles engage in a deep empirical exploration of the conditions under which humans and machines are judged differently for making the same decisions—one in the context of the classic trolley dilemma (Malle et al., 2025), and one in the context of euthanasia decisions (Laakasuo et al., 2025). These two papers, reporting over 20 studies in total, are a striking illustration of the need to leave no stone unturned if we are to reach a satisfying theory of the way people judge machines. Finally, Arnestad and colleagues document a worrying pattern of judgments in the context of autonomous driving: People have a strong preference for human drivers to have an option to regain control manually, but the mere existence of this option partially exonerates machines from blame after a crash, even when regaining control in time is actually impossible (Arnestad et al., 2024).

A final group of three articles engages with an emerging topic in the field of morality and Artificial Intelligence. Up until recently, the field has focused on the morality of machines—but there is an increasing recognition that intelligent machines can af-

fect human morality, through various mechanisms such as enabling new behaviors, exposing humans to new behaviors, or changing the moral value we attach to existing behaviors (Brinkmann et al., 2023; Köbis et al., 2021). Bara and colleagues explore the new moral landscape of AI-generated art, examining the moral stigma attached to the use of machines in artistic pursuit and its impact on aesthetic judgments (Bara et al., 2025). Zhang and colleagues demonstrate how people who are exposed to the unfair behavior of a machine become more desensitized to wrongdoing than when exposed to unfair human behavior—and then show the moral spillover of this experience, namely, a reduction in prosociality (Zhang et al., 2025). Finally, Danaher examines how generative AI, by disrupting the distribution of effective cognitive ability in our societies, may lead us to a reconsider how we define, value, and prioritize equality of opportunity.

Together, these ten papers reveal the values we hope to encode in machines; show how laypeople praise, blame, or empathize with the machines that enact those values; and expose the subtle ways machines feed back into the moral fabric of their creators. We offer this collection as both roadmap and invitation: cognitive scientists need to press further on the questions that will decide whether machines do good, and whether we do good with machines.

Acknowledgments

JFB acknowledges support from grants ANR-17-EURE-0010, ANR-19-PI3A-0004, ANR-22-CE26-0014-01, and ANR-23-IACL-0002.

References

Arnestad, M. N., Meyers, S., Gray, K., & Bigman, Y. E. (2024). The existence of manual mode increases human blame for ai mistakes. *Cognition*, *252*, 105931.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*, 59–64.

Bara, I., Ramsey, R., & Cross, E. S. (2025). Ai contextual information shapes moral and aesthetic judgments of ai-generated visual art. *Cognition*, *257*, 106063.

Bonnefon, J. F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology*, *75*, 653–675.

Brinkmann, L., Baumann, F., Bonnefon, J. F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., & Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, *7*, 1855–1868.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J. F., Brañas-Garza, P., Butera, L., Douglas, K. M., et al. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, *3*, pgae058.

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Köbis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*, 679–685.

Laakasuo, M., Kunnari, A., Francis, K., Košová, M. J., Kopeckỳ, R., Buttazzoni, P., Koverola, M., Palomäki, J., Drosinou, M., & Hannikainen, I. (2025). Moral psychological exploration of the asymmetry effect in ai-assisted euthanasia decisions. *Cognition*, *262*, 106177.

Ladak, A., Loughnan, S., & Wilks, M. (2024). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*, 27–34.

Liu, P., Chu, Y., Zhai, S., Zhang, T., & Awad, E. (2025). Morality on the road: Should machine drivers be more utilitarian than human drivers? *Cognition*, *254*, 106011.

Malle, B. F., Scheutz, M., Cusimano, C., Voiklis, J., Komatsu, T., Thapa, S., & Aladia, S. (2025). People's judgments of humans and robots in a classic moral dilemma. *Cognition*, *254*, 105958.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*, 501–507.

Myers, S., & Everett, J. A. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition*, *256*, 106028.

Purcell, Z. A., Charbit, L., Borst, G., & Nussberger, A.-M. (2025). Estimating divergent moral and diversity preferences between AI builders and AI users. *Cognition*, *263*, 106198.

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., et al. (2019). Machine behaviour. *Nature*, *568*, 477–486.

Reinecke, M. G., Wilks, M., & Bloom, P. (2025). Developmental changes in the perceived moral standing of robots. *Cognition*, *254*, 105983.

Zhang, R. Z., Kyung, E. J., Longoni, C., Cian, L., & Mrkva, K. (2025). Ai-induced indifference: Unfair ai reduces prosociality. *Cognition*, *254*, 105937.