frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 11 March 2024
DOI 10.3389/fpsyg.2024.1382743

OPEN ACCESS

EDITED AND REVIEWED BY
Gerald Matthews,
George Mason University, United States

*CORRESPONDENCE
Feng Yu
✉ psychpedia@whu.edu.cn
Chris Krägeloh
✉ chris.krageloh@aut.ac.nz
Jaishankar Bharatharaj
✉ jaishankar.roboin@gmail.com
Xiaojun Ding
✉ xiaojunding@xjtu.edu.cn

RECEIVED 06 February 2024
ACCEPTED 29 February 2024
PUBLISHED 11 March 2024

CITATION
Yu F, Krägeloh C, Bharatharaj J and Ding X (2024) Editorial: Moral psychology of AI. *Front. Psychol.* 15:1382743. doi: 10.3389/fpsyg.2024.1382743

Editorial: Moral psychology of AI

Feng Yu[1]*, Chris Krägeloh[2]*, Jaishankar Bharatharaj[3]* and Xiaojun Ding[4]*

[1]Department of Psychology, Wuhan University, Wuhan, China, [2]PAIR Lab, Department of Psychology and Neuroscience, Auckland University of Technology, Auckland, New Zealand, [3]PAIR Lab, Bharath Institute of Higher Education and Research, Chennai, India, [4]Department of Philosophy, School of Humanities and Social Sciences, Xi'an Jiaotong University, Xi'an, China

KEYWORDS

moral psychology of AI, moral judgment, robopsychology, mind uploading, uncanny valley, autonomous vehicle, human-robot interaction, AI algorithm

Editorial on the Research Topic
Moral psychology of AI

The recent advancements in Artificial Intelligence (AI) have significantly expanded technological boundaries, prompting an urgent re-assessment of ethical guidelines and AI's role in science in order to address the more complex interactions with these advanced systems (Krägeloh et al., 2022; Ladak et al., 2023). Robopsychology and other emerging fields that provide a nexus of ethics, cognitive science, and AI, critically examine AI's capabilities to perceive, learn, and make decisions that carry moral weight. As AI matures, dissecting its ethical implications is imperative, paralleling the importance of its technical innovations and signaling a pivotal juncture in the discourse on AI (Xu et al., 2022; Zhang et al., 2023; Bonnefon et al., 2024).

The Research Topic "*Moral psychology of AI*" probes the nuanced interplay between ethically-aligned AI and its assimilation into society, evaluating how AI converges with human moral constructs. The studies included in this issue address a range of topics, from the ethical frameworks in autonomous vehicle algorithms and moral perceptions in human-robot interactions, to the fairness of AI in education, the impact of robot aesthetics on moral judgments, and the existential implications of speculative technologies like mind uploading. This editorial synthesizes five important papers that broaden our comprehension of ethical AI, dissecting the intricate relationship between AI and moral principles.

Sui established a benchmark for public preferences in moral algorithms for autonomous vehicles (AVs), probing the ethical frameworks guiding their critical decision-making. Utilizing a survey of 460 Chinese participants about the so-called trolley problem, the study contrasts Utilitarianism, Rawlsianism, Egoism, and a Hybrid model in AV ethics, uncovering a tension between algorithmic preference and purchase intent. While the study shows over half of the respondents' reluctance to purchase AVs equipped with an "egoism" algorithm, it revealed preference for a Hybrid model underscoring the complexity of aligning moral preferences with AI design. This contributes to the moral psychology narrative by identifying ethical priorities for AI applications.

Chen et al. progressed from Sui's foundational work to examine the dynamics of morality and reputation in human-robot interaction, crucial for collaborative potential. Their investigation demonstrates that humans apply moral and reputational considerations to robots, although distinct from human interactions. Through a series of three experiments, the study elucidates how reputation influences the interplay between moral considerations and sharing behavior, varying with the agent's nature—robotic or

human. This research is instrumental for crafting conductive human-robot collaboration, appreciating the influence of moral and reputational perceptions in teamwork dynamics.

Progressing to practical AI applications, Chai et al. investigated AI evaluators in education. The study, with 466 participants, suggests students view AI as fairer than human teachers due to transparency. However, this perception aligns with human evaluators when AI's decisions are elucidated. This links AI ethics with transparency and the necessity for feedback, resonating with the moral preferences and trust issues in earlier studies by Sui and Chen et al.. It also touches upon ethical dilemmas like privacy and algorithmic dependence, emphasizing the delicate equilibrium of AI in educational ethics.

Laakasuo pivoted to the influence of robot aesthetics on human moral judgments. Findings indicate that robots with human-like appearances are treated more leniently for utilitarian actions, while "creepy" robots align better with deontological choices—determined through photorealistic depictions. This challenges prior research and highlights the role of visual design in moral psychology, advocating for consistent imagery in AI ethics studies.

Expanding the moral discourse, Laakasuo et al. explored the moral implications of speculative technologies like mind uploading. The investigation, featuring 1,007 participants, uncovers a correlation between existential beliefs and moral stances on mind uploading, with those valuing existential mattering and afterlife beliefs showing less moral support for the concept. This paper intersects technology and immortality with personal beliefs, suggesting mind uploading prompts a reevaluation of human existence and afterlife, intersecting religious beliefs, death anxiety, and the embrace of AI as a secular promise of immortality.

Collectively, these papers trace a thematic journey from AI's general moral preferences to specific applications (autonomous vehicles, education, human-robot interaction), the impact of physical robot design on moral judgments, and ultimately, the profound existential questions AI poses. This anthology of research intricately ventures into the nuanced interplay between AI and moral reasoning, positing a future where AI not only supplements but also sharpens human ethical judgments. Bridging theoretical discourse with empirical analysis, these publications lay a cornerstone for evolving toward an era where AI systems are intrinsically infused with ethical principles that resonate with societal values and human morality. The overarching aim is to harmonize AI's trajectory with our ethical compass, underscoring the criticality of fostering AI that is developed conscientiously and with ethical clarity.

All in all, this research initiative signals the commencement of a crucial conversation on reframing our ethical paradigms to keep pace with AI's evolution. This compilation will spark continued research and wider discourse on the ethical implications of AI, contemplating the ways in which AI reflects and reshapes our moral fabric, trust dynamics, and societal structures. It highlights the imperative for ethically conscious AI design, exploring themes of public trust, developer accountability, and the broader societal reverberations of AI incorporation. Future scholarly endeavors should seek to inspire ongoing inquiry of AI-related topics in a broader range of relevant fields such as psychology and other disciplines (Krägeloh et al., 2023), aspiring to craft AI that is both transparent and congruent with human values and ethics (Gabriel, 2020; Xu and Yu, 2020).

Author contributions

FY: Writing—original draft. CK: Writing—review & editing. JB: Writing—review & editing. XD: Writing—original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the National Social Science Funds "Research on the moral responsibility attribution of anthropomorphic AI" (Grant No. 20CZX059), "Research on the model of philosophical counseling based on analytic philosophy" (Grant No. 20FZXB047), and the MOE (Ministry of Education in China) Project of Humanities and Social Sciences "Research on epistemic norms of rational actions" (Grant No. 19YJC720006).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bonnefon, J. F., Rahwan, I., and Shariff, A. (2024). The moral psychology of Artificial Intelligence. *Annu. Rev. Psychol.* 75, 653–675. doi: 10.1146/annurev-psych-030123-113559

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds Mach.* 30, 411–437. doi: 10.1007/s11023-020-09539-2

Krägeloh, C. U., Bharatharaj, J., Albo-Canals, J., Hannon, D., and Heerink, M. (2022). The time is ripe for robopsychology. *Front. Psychol.* 13:968382. doi: 10.3389/fpsyg.2022.968382

Krägeloh, C. U., Bharatharaj, J., Heerink, M., Hannon, D., and Albo-Canals, J. (2023). Robots, neurodevelopmental disorders, and psychology: A bibliometric

analysis and a case made for robopsychology. *Adv. Neurodev. Disord.* 7, 290–299. doi: 10.1007/s41252-023-00318-5

Ladak, A., Loughnan, S., and Wilks, M. (2023). The moral psychology of artificial intelligence. *Curr. Dir. Psychol. Sci.* 18:9637214231205866. doi: 10.1177/09637214231205866

Xu, L., and Yu, F. (2020). Factors that influence robot acceptance. *Chin. Sci. Bull.* 65, 496–510. doi: 10.1360/TB-2019-0136

Xu, L., Yu, F., and Peng, K. (2022). Algorithmic discrimination causes less desire for moral punishment than human discrimination. *Acta Psychol. Sinica.* 54, 1076–1092. doi: 10.3724/SP.J.1041.2022.01076

Zhang, Y., Wu, J., Yu, F., and Xu, L. (2023). Moral judgments of human vs. AI agents in moral dilemmas. *Behav. Sci.* 13:181. doi: 10.3390/bs13020181