frontiers in Robotics and AI

TYPE Opinion
PUBLISHED 13 April 2026
DOI 10.3389/frobt.2026.1769361

OPEN ACCESS

EDITED BY
Luisa Damiano,
Università IULM, Italy

REVIEWED BY
Paul Gérard Dumouchel,
Université du Québec à Montréal, Canada
Antonio Fleres,
Università IULM, Italy

*CORRESPONDENCE
Jana Sedlakova,
jana.sedlakova@ethox.ox.ac.uk

RECEIVED 16 December 2025
REVISED 24 March 2026
ACCEPTED 26 March 2026
PUBLISHED 13 April 2026

CITATION
Sedlakova J (2026) What ethical AI and robots could mean: a conceptual and ethical reflection.
Front. Robot. AI 13:1769361.
doi: 10.3389/frobt.2026.1769361

What ethical AI and robots could mean: a conceptual and ethical reflection

Jana Sedlakova*

Nuffield Department of Population Health, Ethox Centre, University of Oxford, Oxford, United Kingdom

KEYWORDS
AI ethics, human-AI collaboration, human-AI interaction, relational ethics, sociotechnical systems, interdisciplinarity

1 Introduction

Recent advances in robotics and artificial intelligence (AI) are reshaping assumptions about moral agency, and moral or ethical interaction. The moral sphere is being expanded to non-human artificial agents. While in history, we expanded the moral sphere to non-human agents such as animals, it is the first time we expand it to something artificially created by us. AI and robots enter and co-shape our epistemic, social, ethical and professional spaces, and create new forms of collaboration and relationships. This raises a fundamental conceptual challenge: before we ask what makes robots or AI ethical or conscious, we must clarify what these terms mean in this novel context, and what assumptions shape the debates and ethical evaluation themselves.

This paper argues that the ethics of AI and robots requires moving beyond human-centered analogies and instead developing a framework that understands AI and robots on their own terms, acknowledges their strengths and limitations, and situates ethics within broader sociotechnical systems of meaning, values, and relationships. While the core arguments apply to both AI systems and robots, a closer ethical analysis requires distinguishing between these two (Küçükuncular, 2026). Within the scope of this opinion piece, I focus primarily on AI systems, while indicating where robots introduce distinct considerations, particularly through embodiment.

2 The need to move beyond the humanization narrative

Much of the contemporary debate, both public and academic, is shaped by a humanization narrative (Salles et al., 2020; Placani, 2024; Sedlakova, 2024b; 2024a). AI systems are described as simulating human cognition, empathy, or reasoning, or even exceeding human abilities and becoming conscious. Questions such as "Is AI conscious?" or "Is AI empathetic?" have a prominent place in discussions of ethical AI. While intuitively appealing, these questions risk obscuring more than they reveal by framing AI primarily through a human lens.

Humanization seems to be a powerful narrative tool. It can help to engage with complex technologies and provide intuitive anchors for ethical concern. However, it also introduces significant systematic challenges that motivate the need to move beyond the humanization narrative. In particular, I focus on three risks: 1) a normative gap between simulated and moral capabilities, 2)

shifts of meaning of key concepts when applied to AI, and 3) methodological problems in evaluating AI when using human-centered benchmarks. Together, these challenges can create misleading expectations about AI's role and its capabilities, and narrow the space of possible AI understandings, its design choices and ways of collaboration.

When AI is described as "listening without judgement," "understanding," or "caring," these descriptions often refer to simulated conversational or behavioral patterns, not to experiential or moral capacities. This creates a normative gap: while AI simulates human abilities like empathy or trust, it cannot fulfil the moral requirements necessarily connected with these, such as being responsible (Sedlakova, 2024b; Sedlakova et al., 2025). While AI could be ethical by aligning its interaction with our values and principles, it is problematic to see it as a moral agent. Moral agency requires, for example, moral understanding, consciousness, responsibility, and lived experiences, which AI lacks (Véliz, 2021). The normative gap can lead to wrong expectations from the human-like AI, for example, perceiving it as having high authority. In clinical or mental health settings, conversational AI communicating in a confident, empathetic, and human-like manner may be perceived as a competent advisor. Users may follow its recommendations even when they are false or inappropriate because the system's interaction style signals understanding and authority that it does not possess.

When we attribute human concepts to AI and interact with it, there is always a change of meaning (Coeckelbergh, 2021; Sedlakova et al., 2025). Trust between a psychotherapist and a client is different from trust between AI and a client. For example, trusting a therapist means that a therapist is responsible for their behavior, can justify their actions and apologize. Can AI apologize? Is it meaningful when it does? AI functions differently from humans (Bender et al., 2021; Felin and Holweg, 2024). AI systems operate within architectures of statistics, prediction, optimization, and classification. Their computational power and strength could be hidden behind the human-like conversational or interaction design which can lead to misplaced ethical evaluation and its focus, for example, on human metrics. Even though AI systems can support ethical outcomes without being moral agents, clarity about their role in relation to ethical requirements is necessary. Recognizing the importance of the distinction between AI systems and humans helps avoid both over-attribution of responsibility to AI systems and under-recognition of human responsibility in design, deployment, and governance. This difference should also be seen as an opportunity for creating novel ways of interaction that might not be enabled if the focus is overly on the simulation of humans. If we rely too much on the humanization narrative, alternative ways of imagining ethical interaction and collaboration remain underexplored. For example, understanding of AI in an interaction and collaboration that emphasizes complementarity rather than imitation, and coordination rather than equivalent capabilities.

Another challenge is evaluating AI based on benchmarks and measures that were developed for human agents and for human-human interaction. As a recent study showed, there is a systematic problem with AI benchmarks due to insufficient definitions, problems with internal validity and statistical rigour (Bean et al., 2025). This gains in difficulty when complex ethical and philosophical or psychological concepts are used for measurements. For example, asking whether AI is conscious or empathetic implicitly assumes that consciousness and empathy are well-defined, measurable human properties that can be scaled or replicated. Yet, we lack a unified account of consciousness even in humans (Ferrante et al., 2025). Treating AI as a candidate for human-like consciousness may therefore reflect more about human projection than about AI itself. What is needed is strong construct validity with clear definitions, representative tasks, and sound statistical analysis (Bean et al., 2025). What is also needed is a well-grounded understanding of AI and its role in our practices.

3 Ethical human-AI and human-robot interaction and collaboration: relational and sociotechnical account

An alternative account of ethical AI and robotics moves beyond evaluating artificial systems in terms of their similarity to humans and instead focuses on the conditions under which human–AI and human–robot interactions and collaborations become ethically acceptable or problematic. From this perspective, it is more meaningful to talk about ethical AI and robots as always embedded within practices of interaction and collaboration with humans. This perspective makes it explicit that AI and robots are always part of our relationships, necessarily shaped by human goals, meanings and values as well as by broader sociotechnical environments (Coeckelbergh, 2021).

What is at stake, then, is not whether AI or robots possess moral or conscious capacities or other human abilities, but whether the forms of interaction and collaboration they enable lead to harm or support beneficial and meaningful collaborations aligned with the normative expectations of the specific domain in which they are deployed. This shift in focus and also methodology allows ethical evaluation to move away from abstract questions about agents' properties toward concrete questions about how relations with AI are constituted, governed, and evaluated in practice. Relational accounts of AI ethics emphasize that ethical meaning does not reside in the intrinsic properties of an agent but emerges through situated interactions, shared practices, normative frameworks, and expectations embedded within sociotechnical systems (Coeckelbergh, 2021; Puzio, 2024; Reinecke et al., 2025). From this standpoint, the central ethical questions focus on how particular ways of relating to AI reshape responsibilities, authority, trust, and norms of collaboration. Importantly, these relations are context-dependent: what counts as ethically appropriate interaction differs substantially across domains, such as healthcare, education, or military applications, each of which carries distinct values, vulnerabilities, and risk profiles.

A recent large-scale meta-analysis shows that combining humans and AI does not automatically lead to better outcomes; rather, performance gains depend critically on how interaction is structured, how tasks are allocated, and how coordination between human and artificial agents is designed (Vaccaro et al., 2024). Human–AI systems often fail to outperform either humans or AI alone when interaction is poorly designed. Another study (Zając et al., 2024) situates core ethical principles such as explainability in the sociotechnical environment, showing how requirements for explainability depend on user expertise level, patient context, medical knowledge and clinical type. McCradden

and Stedman (McCradden and Stedman, 2024) emphasize that the AI explainability in clinical decision-making must be contextualized. They argue that explainability alone is insufficient for good clinical decisions unless it is situated within broader considerations such as the goals of care, the specific patient and social context, and the clinician's responsibility to exercise reasonable professional judgment.

These studies underscore that ethical and effective collaboration with AI emerges from understanding relational configuration. Thus, ethical human–robot and human-AI interaction and collaboration cannot be reduced to isolated technical features or to simulation of human characteristics or abilities. It is key to look at the ways AI systems are embedded in social practices, institutions, and meaning-making processes. Ethical design must therefore attend not only to functionality but to how embodiment shapes relational dynamics and how it might expand vulnerability (Tavory, 2024).

Robots, in addition, introduce the dimension of embodiment, which significantly affects ethical interaction (Coeckelbergh, 2021; Nyholm et al., 2023; Torras, 2024). Physical presence can heighten expectations of agency, responsiveness, emotional presence, and care. An embodied robot does not merely produce outputs; it occupies space, moves among humans, and participates in shared environments. These features give rise to distinct ethical considerations that cannot be fully captured by analyses developed for non-embodied AI systems.

4 Discussion

The arguments developed in this paper imply that moving beyond the humanization narrative requires interdisciplinary work that pays attention to the relational and sociotechnical dimensions in which AI systems and robots are embedded and the conditions under which their use becomes ethically acceptable. Writing from the perspective of an ethicist, I focus on clarifying how ethics contributes to such collaboration by outlining a structured process that integrates several forms of expertise necessary for imagining, designing and evaluating ethical AI systems. Moreover, I discuss recurring disagreements and tensions between disciplines that arise when integrating ethical analysis.

Interdisciplinary work on ethical AI or robots should begin with a specific use case. Ethical evaluation should not be conducted at the level of AI or robots in general, but must consider particular systems, domains, and forms of interaction. A first step is therefore to identify the purpose of the system, the type of interaction or collaboration it is intended to support, and normative expectations together with key values of the domain. For example, in psychotherapy, ethical interaction relies on norms and values such as trust, honesty, responsibility, or care. Understanding these norms provides the context in which the ethical evaluation of AI systems must take place.

The overall strategy should not focus on AI simulation of human abilities, for example, whether AI systems are empathetic, but on how and to what extent AI systems support or hinder these norms. Ethical evaluation thus shifts from focusing on AI properties to evaluating the role AI systems play within existing relations, practices and networks of values.

When AI is designed with a human-like feature, ethical analysis clarifies the purpose of this feature, how it relates to the domain-specific norms and values, and what weaknesses and strengths AI simulation brings. This analysis requires technical expertise regarding AI design and functioning, which also grounds conceptual and ethical analysis in real applications. This step also examines whether alternative approaches beyond humanization could better support the intended purpose. Ethics provides here conceptual and normative clarification, analyzing which concepts remain relevant and in which sense, how their meaning changes when applied to AI systems, what novelties they bring, where their limits lie and what kinds of conceptual adjustments or alternatives are ethically desirable. Such ethical analysis does not focus on one value or one human-like feature, but places them in a network of values and evaluates the trade-offs that arise between them.

Domain expertise becomes particularly important here. It provides experts' knowledge on normative expectations from the application domain as well as empirical understanding of how interactions function within the domain and whether certain forms of interaction benefit users more than others. This helps assess whether other forms of interaction would better support the intended purpose. For example, forms of companionship may be achieved through interaction models inspired by human–animal relations rather than imitation of human therapists. Insights from fields such as biology, ethology, or the humanities can provide models of coordination, communication, and care that do not rely on imitation of human cognition or emotion. By integrating such perspectives, interdisciplinary work moves beyond the question of whether AI resembles humans toward exploring how it might contribute meaningfully within sociotechnical systems.

Even when technical, domain, and ethical expertise are combined, an important dimension remains: AI systems are designed for end users and the public. Their perspectives must therefore be integrated to ensure that ethical evaluation reflects real-world contexts and expectations. Public engagement is particularly important given that expectations are often shaped by human-centered narratives. Engaging users allows these assumptions to be examined and challenged, while ensuring that those affected by AI systems participate in shaping their development.

At the same time, interdisciplinary collaboration is often challenging. Differences in methods, vocabularies, and epistemic standards can lead to misunderstandings or disagreements between disciplines. Ethics, in particular, is sometimes perceived as too abstract, impractical, or disconnected from real-world applications. Conversely, ethical analysis may question assumptions that are taken for granted in technical design processes or empirical domains.

Several types of disagreement commonly arise in interdisciplinary work. One concerns the role of ethical concepts. Technical research may treat concepts such as fairness, transparency, or trust as measurable properties of systems, while ethical analysis emphasizes that these concepts are embedded in social practices and normative expectations. A second source of disagreement concerns the scope of evaluation. Technical approaches may focus on optimizing specific system features, whereas ethical analysis may highlight broader sociotechnical contexts, including institutional responsibilities, power relations, and long-term societal consequences. A third tension concerns the role of empirical and normative aspects and how they relate to the ethical evaluation.

When empirical studies show that there are positive outcomes associated with AI simulating empathy, it does not mean that such simulation is ethically good from normative reasons. This can be illustrated by our societal and ethical attitude towards lying. There could be empirical studies showing that lying pays off, but this does not mean that lying is ethically good.

Addressing these tensions requires practices that enable meaningful collaboration across disciplinary boundaries. While technical and domain expertise help make ethical analysis more grounded and practically relevant, ethics, in turn, contributes by clarifying key concepts, identifying normative expectations, and making explicit the value assumptions that shape AI development and deployment. Approaches such as adversarial collaboration (Kahneman and Klein, 2009; Nature, 2025) and adversarial co-operation (Parker, 2026) provide useful models by treating disagreement as a starting point for joint inquiry.

In this context, disciplines work together to articulate their assumptions, clarify points of disagreement, and jointly examine how system capabilities, domain practices, and normative expectations interact in specific use cases. Tools such as visualizations can support this process by making system behavior, data flows, and value trade-offs visible and open to shared analysis. Rather than forcing consensus, the aim is to make differences in concepts, methods, and normative commitments constructive. In this way, disagreement becomes a resource for refining concepts, identifying hidden assumptions, and exploring alternative design choices.

Importantly, ethical human–AI and human-robot interaction and collaboration should be understood as an evolving process rather than a static checklist. As AI systems enter new domains and reshape social practices and meaning, both technical capabilities and normative expectations change. Sustaining ethical alignment, therefore, requires ongoing interdisciplinary dialogue and reflexivity, rather than one-off ethical assessments. This process demands epistemic humility: recognition of uncertainty, openness to revision, and respect for the limits of any single disciplinary perspective.

Author contributions

JS: Conceptualization, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The research for this paper was undertaken as part of the ANTITHESES Discovery Research Platform for Transformative Inclusivity in Ethics and Humanities Research. ANTITHESES is funded by a grant from the Wellcome Trust 226801/Z/22/Z.

Acknowledgements

The author acknowledges the use of an AI-based tool for assistance with language editing and stylistic refinement.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI tools were used for language editing and stylistic refinements.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., et al. (2025). Measuring what matters: construct validity in large language model benchmarks. *arXiv*. doi:10.48550/arXiv.2511.04703

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Virtual event, Canada: association for computing machinery*. doi:10.1145/3442188.3445922

Coeckelbergh, M. (2021). Three responses to anthropomorphism in social robotics: towards a critical, relational, and hermeneutic approach. *Int. J. Soc. Robotics* 14, 2049–2061. doi:10.1007/s12369-021-00770-0

Felin, T., and Holweg, M. (2024). Theory is all you need: AI, human cognition, and decision making. *Strategy Sci.* 9 (4), 297–514. doi:10.1287/stsc.2024.0189

Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature* 642 (8066), 133–142. doi:10.1038/s41586-025-08888-1

Kahneman, D., and Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64 (6), 515–526. doi:10.1037/a0016755

Küçükuncular, A. (2026). Robots and AI are not one moral category: why the distinction matters for ethical and conscious systems. *Front. Robotics AI* 13, 1776097. doi:10.3389/frobt.2026.1776097

McCradden, M. D., and Stedman, I. (2024). Explaining decisions without explainability? Artificial intelligence and medicolegal accountability. *Future Healthc. J.* 11 (3), 100171. doi:10.1016/j.fhj.2024.100171

Nature (2025). "Make science more collegial: why the time for "adversarial collaboration" has come", 641(8062), pp. 281–282. doi:10.1038/d41586-025-01379-3

Nyholm, S., Friedman, C., Dale, M., Puzio, A., Babushkina, D., Löhr, G., et al. (2023). *Social robots and society*, 53–82.

Parker, M. J. (2026). Bioethics and the value of disagreement. *J. Med. Ethics* 52 (1), 7–13. doi:10.1136/jme-2024-110174

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI Ethics* 4 (3), 691–698. doi:10.1007/s43681-024-00419-4

Puzio, A. (2024). Not relational enough? Towards an eco-relational approach in robot ethics. *Philosophy and Technol.* 37 (2), 45. doi:10.1007/s13347-024-00730-2

Reinecke, M. G., Kappes, A., Porsdam Mann, S., Savulescu, J., and Earp, B. D. (2025). The need for an empirical research program regarding human–AI relational norms. *AI Ethics* 5 (1), 71–80. doi:10.1007/s43681-024-00631-2

Salles, A., Evers, K., and Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience* 11 (2), 88–95. doi:10.1080/21507740.2020.1740350

Sedlakova, J. (2024a). Conversational AI for psychotherapy and its role in the space of reason. *Cosmos+ Taxis* 12 (5+ 6), 80–87.

Sedlakova, J. (2024b). *Ethical and epistemic challenges of conversational AI in mental healthcare: interdisciplinary inquiry for responsible human-AI interaction*. Dissertation. Zurich: University of Zurich. doi:10.5167/uzh-269480

Sedlakova, J., Lucivero, F., Pavarini, G., and Kerasidou, A. (2025). Human-like epistemic trust? A conceptual and normative analysis of conversational AI in mental healthcare. *Am. J. Bioeth.* 0 (0), 1–16. doi:10.1080/15265161.2025.2526734

Tavory, T. (2024). Regulating AI in mental health: ethics of care perspective. *JMIR Mental Health* 11, e58493. doi:10.2196/58493

Torras, C. (2024). Ethics of social robotics: individual and societal concerns and opportunities. *Annu. Rev. Control, Robotics, Aut. Syst.* 7, 1–18. doi:10.1146/annurev-control-062023-082238

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* 8 (12), 2293–2303. doi:10.1038/s41562-024-02024-1

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI and Soc.* 36 (2), 487–497. doi:10.1007/s00146-021-01189-x

Zając, H. D., Ribeiro, J. M. N., Ingala, S., Gentile, S., Wanjohi, R., Gitau, S. N., et al. (2024). ""it depends": configuring AI to improve clinical usefulness across contexts," in *Proceedings of the 2024 ACM designing interactive systems conference* (New York, NY, USA: Association for Computing Machinery DIS '24), 874–889. doi:10.1145/3643834.3660707