Can Interactive Mediation Serve as a Method in Robot Ethics?

Zhengqiang Han

University of Chinese Academy of Sciences, Beijing, China

hanzhengqiang17@mails.ucas.ac.cn

Abstract. Existing research in robot ethics has focused primarily on norm embedding and moral learning as ways to align robotic behavior with human values, while paying little attention to the normative role of artifacts in robotic action. Drawing on technological mediation theory, this article examines whether and how artifacts can mediate robots' moral action. It argues that although robots lack human forms of agency, their perception–action cycles provide the *interactive conditions* under which technological mediation can occur. On this basis, the article distinguishes between *mediation of experience* and *mediation of praxis* in robotic contexts and introduces the concept of *interactive mediation* to explain how artifacts structurally reshape robots' spaces of possible action. It thus shows that technological mediation can serve as a complementary approach in robot ethics, distinct from norm embedding and moral learning.

Keywords: technological mediation; robot ethics; interactive mediation

1 Introduction

Current research on enabling robots to act in morally appropriate ways has primarily followed two paths. One is a top-down approach that embeds moral norms into robotic systems so that decisions are guided by explicit rules[1,2]. The other is a bottom-up approach that equips robots with moral learning capacities, allowing appropriate behavioral patterns to emerge from experience[3,4]. Both approaches face structural difficulties: the former struggles to formalize highly contextual human moral norms into executable rules, while the latter risks reinforcing undesirable behavioral patterns in complex environments, thereby undermining moral reliability.

These difficulties suggest that robot ethics concerns not only normative content or agent capacities, but also the action environment itself. Robots do not operate in abstract spaces; rather, they are always embedded in material–symbolic environments composed of artifacts, institutions, and technical infrastructures. Yet most work in robot ethics continues to treat the environment as a neutral background rather than as an active factor in moral shaping.

The theory of technological mediation offers an important perspective for rethinking this issue. It holds that artifacts are not value-neutral tools but intervene deeply in moral

S. Garcia-Esteban et al. (eds.), *Proceedings of the 2026 5th International Conference on Social Sciences and Humanities and Arts (SSHA 2026)*, Advances in Social Science, Education and Humanities Research 1014,
https://doi.org/10.2991/978-2-38476-577-5_105

action by shaping experience and practice[5,6]. If artifacts can mediate human perception and action, a question arises that has not yet been systematically addressed: can artifacts also mediate robot action at the environmental level, enabling robots to conform more reliably to human value norms?

Accordingly, this article poses the following research questions: in the absence of human-like agency structures, can technological mediation still occur in robotic systems? If so, through what mechanisms, and how can these acquire normative significance within robot ethics? To address these questions, this article brings technological mediation theory and artifact theory into the robotics context, analyzes the conditions under which robots can become objects of mediation, and proposes the concept of interactive mediation, aiming to provide an environment-oriented supplementary pathway for robot ethics.

2 Robotic Action from the Perspective of Technological Mediation

Technological mediation theory holds that technology is not a neutral medium but plays a constitutive role in shaping relations between subjects and the world. Its core mechanisms are *mediation of experience* and *mediation of praxis* [5,7]. The former concerns how artifacts reshape structures of perception and understanding, while the latter concerns how artifacts guide and constrain action through "scripts" [6]. A microscope, for instance, alters what counts as observable[8], while a speed bump translates the norm of "safety first" into a material constraint.

These studies show that moral practice is not grounded solely in subjective judgment but is embedded in sociotechnical structures composed of artifacts. Yet existing formulations almost exclusively take humans as objects of mediation, treating free will, intentionality, and normative understanding as operative presuppositions. As robots increasingly enter social practice, a key question arises: must technological mediation be confined to human subjects? If robots lack free will and intentionality, can they nevertheless become objects of mediation?

This article argues that denying this possibility would prematurely restrict the explanatory scope of technological mediation theory. Although robots do not possess human-like agency structures, their actions are always embedded in technological environments composed of artifacts, institutions, and algorithms, and they continuously interact with these elements through *perception–action cycles*. This suggests the need to reconsider the conditions under which mediation occurs and to explore its possible forms in non-human action systems.

It is important to distinguish between robots as mediators and robots as objects of mediation. This article focuses on the latter, namely, how artifacts shape robotic perception and action through their material and symbolic structures. Extending technological mediation theory to robotics thus does not require replicating models of human agency, but rather redefining the structural conditions under which mediation occurs. The next section combines artificial agency theory with the perception–action cycle

framework to argue that robots satisfy the *interactive conditions* for technological mediation.

3 Interactive Conditions: Robots as Objects of Technological Mediation

In technological mediation theory, humans are typically regarded as objects of mediation on the basis of two core conditions of agency: free will and intentionality. The former refers to the capacity to initiate and regulate action autonomously [7], while the latter refers to the purposive structure by which action is directed from internal mental states toward external goals[8]. Together, these constitute the philosophical foundations of the traditional notion of a "moral agent."

Contemporary robotic systems clearly do not satisfy these conditions. Their action spaces are determined by algorithmic architectures, sensor inputs, and objective functions, and their "choices" result from variable mappings and parameter updates rather than from intentional states or value reflection[9]. Accordingly, treating robots as moral subjects equivalent to humans lacks both metaphysical justification and engineering feasibility[10-11].

This does not mean, however, that robots cannot participate in ethical practice. Rather, it is necessary to distinguish between *moral agents* and *moral entities* and to reconsider the minimal structural conditions under which technological mediation can occur. Floridi and Sanders's (2004) [12]framework of *artificial agency* understands agency as a system-level functional structure rather than a psychological property, holding that any system exhibiting interactivity, autonomy, and adaptability can count as an agent in an operational sense. On this view, robots are not moral agents but *moral entities*, insofar as the outcomes of their actions are open to ethical evaluation[1].

From an engineering perspective, contemporary robotic systems already possess stable mechanisms for translating environmental signals into action constraints, for instance by mapping traffic signs, spatial boundaries, and normative cues into path-planning and control parameters through multimodal perception and control algorithms[13]. Although this process involves no normative understanding, it achieves a functional execution of value structures.

Yet the artificial agency framework alone does not explain how robots become objects of technological mediation. To address this gap, this article introduces the *perception–action cycle* as the structural condition of mediation: systems continuously perceive their environment, generate actions, and update their states through feedback, forming stable yet plastic interaction structures[14,15,16]. Within this cycle, artifacts enter perceptual channels through their material or symbolic structures and are mapped at the algorithmic level into normative constraints or action parameters, thereby shaping robots' spaces of possible action. Technological mediation thus operates not through "understanding–judgment–choice," but through a structural mechanism of "perception–mapping–execution."

On this basis, the minimal conditions under which robots can become objects of technological mediation are summarized as *interactive conditions*: any system that

possesses a stable perception–action cycle and can translate artifact structures into action constraints can be an object of mediation, even in the absence of human-like free will or intentionality. This shows that technological mediation can be grounded in interactive system architectures, offering a third path for robot ethics distinct from normative embedding and moral learning.

4 Interactive Mediation: How Artifacts Regulate Robotic Action

To illustrate how artifacts operate as mediators in robotic contexts, this section takes the *speed bump* as a paradigmatic case. In technological mediation theory, the speed bump is a classic example of a *scripted norm*, whose script can be summarized as: "Before reaching me, slow down." This script translates the normative requirement of "safety first" into material structures or symbolic cues that shape patterns of action.

In robotic contexts, speed bumps need not take only physical form; they may also appear as symbolic signs or digital signals. Physical speed bumps force mobile robots to decelerate through structural protrusions; symbolic speed bumps prompt slowing through visual patterns; and virtual speed bumps transmit normative constraints via high-definition maps, vehicle–infrastructure communication, or environmental tags. In robotic systems, this mechanism is especially salient: artifacts no longer influence judgment merely through symbolic meaning but enter directly into algorithmic perception and control architectures as components of action constraints.

Mediation of experience concerns how artifacts shape an agent's modes of perceiving the environment. In robotic systems, this occurs when artifacts enter perceptual channels through their material or symbolic features and are mapped by algorithmic modules into normatively salient environmental elements. For example, when a robot detects a speed bump through vision, touch, or map interfaces, its perceptual system not only classifies it as an obstacle or road structure but also tags it as a normatively relevant situation requiring deceleration[17]. By reconfiguring environmental representations, this process renders certain action paths salient while making others inappropriate or unavailable.

In this sense, artifacts convey not only physical information but also normative information. The ethical function of experiential mediation does not lie in endowing robots with moral judgment, but in reshaping their perceivable action spaces so that system-level attention is preferentially directed toward situations involving safety, responsibility, or fairness.

Mediation of praxis focuses on how artifacts directly shape action paths through their scripts. In robotic systems, this occurs when perceptual outputs are mapped into control parameters, path constraints, or priority weights, thereby restructuring the space of possible actions. In the case of a speed bump, once its structure or digital tag is detected, the control system automatically adjusts speed parameters, replans routes, or activates safety modes. This form of mediation is not advisory but structural: by entering the control loop, artifacts suppress certain options (e.g., high-speed traversal) and invite

others (e.g., slowing down or rerouting). In this sense, speed bumps do not merely express norms but enact them.

Unlike humans, robots respond to scripts with high reliability: insofar as artifact structures are robustly perceived, their normative effects are almost invariably translated into behavioral outcomes[18]. In robotic contexts, experiential mediation and practical mediation are not separate processes but form a continuous and closed mediating structure within the *perception–action cycle*. Empirical studies of robot behavior in ethically relevant scenarios highlight that robotic systems exhibit patterned responses to normative cues that may reflect embedded algorithmic behavior more than autonomous moral judgment[19].

This artifact-centered, perception–action-cycle-based mechanism of ethical regulation is termed *interactive mediation*. Whereas traditional technological mediation theory emphasizes how artifacts shape human understanding and choice, interactive mediation highlights how artifacts directly reshape robots' spaces of possible action through structural mapping mechanisms. Here, ethics is expressed less as subjective judgment than as system-level behavioral constraint. The defining feature of interactive mediation is that it does not require robots to possess human-like moral agency or capacities for value reflection, yet it can still realize stable, verifiable, and scalable forms of normative consistency at the system level.

5 Possible Objections and Responses

Two main objections arise concerning the applicability of technological mediation in robotic contexts. The first concerns *possibility*: whether differences in agency structures between humans and robots prevent robots from becoming objects of technological mediation. The second concerns *effectiveness*: whether artifacts can genuinely carry human moral norms, thereby rendering technological mediation ethically ineffective.

Regarding the possibility objection, the key point is not to deny the ontological differences between humans and robots, but to clarify that human agency is not a necessary condition for technological mediation. Although traditional mediation theory centers on human subjects, the mechanisms it identifies do not essentially depend on free will or intentionality. As long as a system can transform environmental features into action constraints through stable structures, technological mediation can occur. While robots lack human autonomy and intentionality, they possess perception and action capacities and can systematically map artifact features into behavioral parameters through algorithmic structures. More fundamentally, the ongoing interaction between robots and artifacts within the *perception–action cycle* itself constitutes the structural condition for mediation—precisely what this article conceptualizes as *interactive conditions*.

Regarding the effectiveness objection—that artifacts cannot carry genuine moral norms—the response is that the normative structures embedded in artifacts exhibit functional equivalence to human moral norms in practice. As Hurshman (2023) [20]argues, artifacts and institutions alike stabilize behavior through structural arrangements, providing theoretical support for embedding norms in material or symbolic forms. In robot ethics, whether norms are embedded in robotic systems themselves or in

environmental artifacts, their practical function is to translate value requirements into action constraints. In this sense, artifacts can be understood as materialized extensions of human moral norms.

Finally, it is important to distinguish between the *possibility* and the *effectiveness* of technological mediation. The former concerns whether robots structurally qualify as objects of mediation; the latter concerns whether specific mediating designs can reliably achieve intended ethical outcomes in practice. This article argues for possibility under the idealized assumption that artifacts can structurally represent value norms, showing that even without human-like agency, robots can be mediated under interactive conditions. At the same time, it acknowledges that effectiveness depends on engineering implementation, contextual adaptation, and modes of normative embedding—issues that constitute central directions for future research and design.

6 Conclusion

This article offers a qualified affirmative answer to the question "Can interactive mediation serve as a method in robot ethics?" By extending technological mediation theory beyond human subjects, it shows that although robots lack human forms of agency, their *perception–action cycles* provide *interactive conditions* under which artifacts can reshape robotic perception and action through structural mapping mechanisms. On this basis, it introduces the concept of *interactive mediation* and distinguishes between *mediation of experience* and *mediation of praxis* in robotic contexts, demonstrating how artifacts can both reconfigure robots' environmental representations and directly shape their action pathways through scripts. Interactive mediation thus emerges as a supplementary approach to robot ethics, distinct from norm embedding and moral learning, while also laying a conceptual foundation for future interdisciplinary research on artifact design, algorithmic interfaces, and the embedding of social values in socio-technical systems.

References

1. Reiter, P., Norman, U., Weinberger, N., & Bruno, B. (2025, July). Artificial Moral Agents: Should Machines Take Ethical Responsibility?. In *2025 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO)* (pp. 218-224). IEEE.
2. Yaacov, D.-D. (2025). Normative moral pluralism for AI: A framework for deliberation in complex moral contexts. arXiv. https://doi.org/10.48550/arXiv.2508.08333
3. González Barman, K., Lohse, S., & de Regt, H. W. (2025). Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives? Philosophy & Technology.
4. Vishwanath, A., Dennis, L., & Slavkovik, M. (2024). Reinforcement learning and machine ethics: A survey. arXiv. https://doi.org/10.48550/arXiv.2407.02425
5. Verbeek, P.-P. (2005). What things do: philosophical reflections on technology, agency, and design. Penn State Press.
6. Latour B. Where are the missing masses? The sociology of a few mundane artifacts[J].Shaping technology/building society: Studies in sociotechnical change,1992.1:225-258.

7. Verbeek, P. P. (2015). Cover story: Beyond interaction: A short introduction to mediation theory. Interactions, 22(3), 26–31.
8. Ihde D (1990) Technology and the lifeworld: from garden to earth. Indiana University Press, Bloomington
9. Frankfurt, H. G. . (1971). Freedom of the will and the concept of a person. Journal of Philosophy, 68(1), 5-20.
10. Bratman, M. . (1987). Intention, plans, and practical reason. Harvard University Press.
11. Maure, R., & Bruno, B. (2025). Autonomy in socially assistive robotics: a systematic review. *Frontiers in Robotics and AI*, *12*, 1586473.
12. Grau, C. (2011). There Is No "I" in "Robot": Robots and Utilitarianism. In M. Anderson & S. L. Anderson (Eds.), Machine Ethics (pp. 451–463). chapter, Cambridge: Cambridge University Press.
13. Selinger, E. Confronting the Moral Dimensions of Technology Through Mediation Theory. Philos. Technol. 27, 287–313 (2014).
14. Floridi, L., Sanders, J. On the Morality of Artificial Agents. Minds and Machines 14, 349–379 (2004).
15. Roychoudhury, A., Khorshidi, S., Agrawal, S. et al. Perception for Humanoid Robots. Curr Robot Rep 4, 127–140 (2023).
16. Domínguez-Vidal, J. E., Rodríguez, N., & Sanfeliu, A. (2025). Perception–intention–action cycle in human–robot collaborative tasks: The collaborative lightweight object transportation use-case. *International Journal of Social Robotics*, *17*(10), 1927-1956.
17. Laina, S. B., Boche, S., Papatheodorou, S., Schaefer, S., & Jung, J. (2025). FindAnything: Open-Vocabulary and Object-Centric Mapping for Robot Exploration in Any Environment. arXiv. https://doi.org/10.48550/arXiv.2504.08603
18. Nguyen, P., Verdoja, F., & Kyrki, V. (2025). Event-Grounding Graph: Unified Spatio-Temporal Scene Graph from Robotic Observations. arXiv. https://doi.org/10.48550/arXiv.2510.18697
19. Söderlund, M. (2023). Service robots and artificial morality: an examination of robot behavior that violates human privacy. *Journal of Service Theory and Practice*, *33*(7), 52-72.
20. Hurshman, C. Artifacts and intervention: a persistence theory of artifact functions. Synthese 202, 128 (2023).