Available Online: 28 June 2026 DOI: 10.54254/2753-7080/2026.34899

# A critique of the technological rationality perspective in practical machine ethics

***Huiyun Guo***

Research Center for Science and Technology Ethics Governance, Southwest University, Chongqing, China

416409803@qq.com

**Abstract.** Within the framework of technological rationality, practical machine ethics takes the legitimacy of autonomous intelligent systems for granted and defines the ethical decision-making dilemmas faced by such systems as problems in technology—namely, issues of ethical safety. Consequently, it adopts the technical strategy of developing moral machines to address these challenges. Building upon a critical examination of the legitimacy of autonomous intelligent systems, this paper argues that the ethical decision-making dilemmas associated with autonomous intelligent systems should instead be understood as problems of technology itself. More specifically, they constitute ethically significant problems of control arising from the autonomy of autonomous intelligent systems. Through a critical analysis of the technocratic orientation of practical machine ethics, this study contends that, while grounded in a conservative instrumental view of technology, practical machine ethics advocates a radical technological solution—the moral machine. The paper ultimately argues that, because autonomous intelligent systems inherently embody control-related problems, this radical technological strategy will inevitably rebel against the very logic of technological rationality.

**Keywords:** practical machine ethics, autonomous intelligent systems, moral machine, technological rationality

## 1. Introduction

With the rapid deployment of generative artificial intelligence technologies such as large language models and autonomous agents, machine ethics has become a central branch of AI ethics research. According to differences in research orientation, scholars generally divide machine ethics into two major approaches: practical machine ethics and philosophical machine ethics [1][1]. Philosophical machine ethics focuses on foundational theoretical issues, including the ontological status and ethical legitimacy of artificial moral agents. Practical machine ethics, by contrast, operates within a framework of technological rationality[2]

[1] Steve Torrance has observed this division in the existing literature and distinguishes between two approaches to machine ethics: "practical machine ethics" (referred to in this paper simply as practical machine ethics) and "philosophical machine ethics".

[2] In this paper, technological rationality refers to a mode of thinking in the philosophy of technology that seeks to solve all problems—including both technical and non-technical problems—through technological means.

and takes the engineering development of "moral machines" as its primary solution, seeking to mitigate the decision-making risks of autonomous intelligent systems by embedding ethical rules into technological systems. This approach has now become the dominant implementation strategy within the engineering community.

The entire framework of practical machine ethics rests upon two largely unexamined assumptions. First, it assumes that the pursuit of ever-greater autonomy in autonomous intelligent systems is an irreversible technological inevitability. Second, it categorizes the ethical conflicts generated by intelligent systems as "problems in technology"—that is, as mere ethical safety deficiencies that can be remedied through further technological design. Representative scholars of this approach, including Michael Anderson, Susan Leigh Anderson, Colin Allen, and Wendell Wallach, have continuously advanced value-sensitive design methodologies and artificial moral reasoning modules, thereby constructing a comprehensive technocratic solution to ethical challenges.

This paper offers a critical examination of the underlying framework of technological rationality that informs practical machine ethics. Rather than following engineering-oriented studies that presuppose the legitimacy of autonomous intelligent system development and seek merely to optimize moral-machine design, this study adopts a fundamentally critical perspective. First, it distinguishes between "problems in technology" and "problems of technology", redefining the conflicts generated by the autonomy of autonomous intelligent systems as deeply ethical issues of human–machine control rather than superficial safety malfunctions. Second, it systematically deconstructs the three core assumptions upon which practical machine ethics relies within its framework of technological rationality and subjects each to theoretical falsification. Third, by contrasting passive restraint-oriented safety technologies with the more radical paradigm of moral machines, it demonstrates that moral machines may further intensify the tension between artificial agency and human moral agency. Finally, the paper reveals a fundamental logical paradox in the proposal to address the ethical dilemmas of autonomous intelligent systems through moral machines: the radical strategy of technological intervention ultimately undermines the central aspiration of technological rationality—namely, human-centered control.

On the one hand, this study fills a gap in critical scholarship at the intersection of machine ethics and technological rationality, enriching theoretical discussions concerning the agency of autonomous artifacts and the allocation of rights and responsibilities between humans and machines. On the other hand, it provides a critical point of reference for the governance of contemporary generative AI and autonomous agents, cautioning scholars against relying exclusively on the technocratic assumption that "technological ethical problems can be solved through technology itself". In doing so, it offers theoretical support for the development of a pluralistic governance framework that balances technological constraints, social regulation, and the cultivation of human moral capacities.

## 2. The technological rationality perspective of practical machine ethics

Horkheimer defines technological rationality as follows: "It is essentially concerned with means and ends, with the suitability of procedures for purposes that are more or less taken for granted or assumed to be self-evident, while rarely questioning whether the ends themselves are reasonable" [2]. Practical machine ethics is grounded precisely in this framework of technological rationality. It presupposes the purposiveness and legitimacy of autonomous intelligent systems, defines the challenges faced by such systems as technical problems, and, at the level of instrumental value, proposes the development of moral machines as a technological solution.

## 2.1. The presupposition of the legitimacy of autonomous intelligent systems

### *2.1.1. The development-trend thesis*

From the perspective of technological development, the advancement of autonomous intelligent systems is regarded not as a matter of human choice but as participation in an uncontrollable technological tide. According to Colin Allen et al.: "The computer revolution continues to drive dependence on automation, and autonomous systems are coming whether we like it or not… Human beings have always adapted to their technological artifacts, and the benefits of autonomous intelligent systems are likely to outweigh their costs. The development of autonomous intelligent systems is an unstoppable technological trend. We must adapt to it and improve it rather than restrict or prevent it" [2].

### *2.1.2. Autonomy as an internal norm of artificial intelligence technology*

Practical machine ethics regards autonomy as an intrinsic norm of artificial intelligence technology. Michael Anderson and Susan Leigh Anderson argue that: "Human micromanagement of autonomous intelligent machines would inevitably sacrifice their capacity for independent operation.… Ideally, we would like to trust autonomous machines to make the correct moral decisions on their own, which requires that we create a moral framework for them" [2].

The pursuit of higher levels of autonomy functions as an implicit internal norm of artificial intelligence technology. In the development of AI, automation is viewed as inherently desirable: the more autonomous a system is, the better; full autonomy represents the ideal technological aspiration. This orientation is clearly reflected in contemporary technological practice. Autonomous driving systems and autonomous weapons are both frontier areas of AI research and development. Their status as cutting-edge technologies signifies the proactive pursuit of greater autonomy, and autonomy itself has become a marker of technological progress in intelligent systems.

## 2.2. Ethical safety problems within the framework of "technical problems"

In the literature of practical machine ethics, one frequently encounters statements emphasizing the urgency of developing moral machines: "We urgently need a functional moral reasoning system because artificial intelligence systems will be deployed on a large scale" [3]. Similarly, "Given their capacity to learn and process information, and their increasing role in autonomous decision-making, existing artificial agents require the immediate and urgent development of computational morality" [4].

These statements reveal a common assumption: as artificial intelligence becomes increasingly autonomous and intelligent, its applications spread into nearly every aspect of social life. As its decisions acquire greater ethical significance, it becomes necessary to ensure that such systems possess "ethical safeguards" [2].

Practical machine ethics recognizes that autonomous intelligent systems face ethical decision-making dilemmas, yet it interprets these dilemmas primarily as a lack of ethical safeguards. Within computer ethics, ethical value concerns are often transformed into technical design issues through value-sensitive design and are therefore treated as technical problems[3]. According to Wang Dazhou's definition, "a technical problem refers to a problem within technology ("problems in technology" or "technological problems"), namely,

[3] The term "problems in technology" is used in contrast to "problems of technology". The former adopts the perspective that autonomous intelligent systems are inherently legitimate and focuses on problems arising within such systems, whereas the latter takes a critical perspective on autonomous intelligent systems themselves. Technological rationality tends to define problems in a reductionist manner, ultimately treating all problems as technical problems. For this distinction, see the article "*Philosophy of Technology, Technological Practice, and Technological Rationality*" by Wang Dazhou and Guan Shixu.

a problem that engineers believe can be solved through technical means" [5]. Although practical machine ethics claims to differ from computer ethics, it nevertheless inherits the technological-rationalist mode of thinking characteristic of computer ethics. Taking the legitimacy of autonomous intelligent systems as a given, practical machine ethics interprets the ethical decision-making dilemmas they encounter as technical problems of the systems themselves—specifically, problems of ethical safety.

### 2.3. The moral-machine safety project under a technocratic framework

Following the technocratic logic of technological rationality, practical machine ethics seeks to resolve the ethical safety problems of autonomous intelligent systems through further technological design. In engineering practice, safety concerns are commonly addressed by translating safety values into concrete technical standards. For example, in order to prevent explosions, the safety value associated with pressure vessels is transformed into engineering specifications such as a required wall thickness. In a similar fashion, scholars such as Michael Anderson and Susan Leigh Anderson [6], Colin Allen and Wendell Wallach [7] extend the logic of value-sensitive design by proposing that ethical dimensions be integrated directly into autonomous intelligent systems. They tend to regard the moral machine as a distinctive safety-engineering project and formulate it as a specific engineering challenge: "How can one design a multi-objective system capable of responding flexibly and autonomously within real-world environments independent of its designers and users" [2].

Roman V. Yampolskiy has explicitly recognized this orientation within practical machine ethics: "Over the past decade, the new subfield of computer science known as machine ethics has flourished. This technology represents only one proposed solution to the safety problem posed by advanced intelligent machines" [8].

For engineers committed to the pursuit of autonomous intelligent technologies, perhaps the most attractive option is to endow autonomous intelligent systems with ethical agency. Such a solution promises to achieve both complete autonomy and conformity with the safety requirements that govern standard engineering practice. From the perspective of practical machine ethics, the development of moral machines therefore appears to be an ideal and highly desirable engineering project. Michael Anderson and Susan Leigh Anderson maintain that: "Rather than viewing intelligent robots as a terrifying threat and seeking to suppress their development, we propose a feasible and more realistic solution" [7]. In this view, a moral robot is ultimately a safer robot.

## 3. Critiquing the presupposed legitimacy of autonomous intelligent systems

### 3.1. A critique of the development-trend thesis

Practical machine ethics maintains that autonomous intelligent systems represent an irresistible trend in technological development. This position is a typical expression of technological determinism. According to the theory of technological autonomy advanced by Jacques Ellul, human beings are determined by technology; human choices and destinies are shaped by the trajectory of technological development.

From the standpoint of technological autonomy, however, it is inherently contradictory to argue for the legitimacy of a specific technological undertaking such as the development of moral machines. If technological autonomy is taken as a justification for developing moral machines, then it can only provide an abstract rationale detached from human agency. Moreover, in practice, societies have already demonstrated considerable caution toward the development of autonomous intelligent systems. Rather than merely adapting passively to technological change, governments and institutions actively seek to regulate and constrain both the pace and scope of autonomous-system development.

From a theoretical perspective, Ellul's conception of technological autonomy has been subjected to substantial criticism. Andrew Feenberg [9] argues that technological autonomy is an essentialist position akin to Martin Heidegger's notion of *Gestell* ("enframing"). Feenberg advances a social constructivist account of technology: "(1) Technological development is determined jointly by progressive technical standards and social standards; consequently, technology can develop along many different trajectories depending upon the dominant social hegemony. (2) When social institutions adapt to technological development, the process of adaptation is reciprocal: technology changes in response to social conditions, while simultaneously acquiring the capacity to influence those conditions" [9]. Similarly, Blay Whitby argues that the development of artificial intelligence is not inevitable and that particular technological pathways can, in principle, be prohibited or restricted [10].

## 3.2. A critique of autonomy as the internal norm of artificial intelligence

Treating the degree of autonomy as the default objective of artificial intelligence technology is a narrow and contestable position. As a technological standard, autonomy must be evaluated within the broader framework of human values. Why should humanity pursue ever-higher levels of autonomy in intelligent systems without limit?

Advocates of machine ethics respond to this question as follows: "There are many tasks that human beings either cannot perform (because they are too dangerous), do not wish to perform, or do not perform as well as machines. In such cases, machines can complete these tasks independently" [2]. This argument suggests that certain categories of work—those that are dangerous, undesirable, or beyond human competence—should be delegated to machines capable of independent operation.

First, the degree of machine independence is a critical issue in such scenarios. Human beings may reasonably choose to delegate tasks to autonomous intelligent systems in pursuit of safety, efficiency, or convenience. Yet this does not necessarily imply that human involvement should be eliminated altogether. Automation does not require the complete replacement of human beings; rather, it can serve to enhance human effectiveness. Why, then, should remote control or supervisory intervention be excluded? The Mars rovers operated by NASA and many deep-sea exploration robots function through remote control mechanisms. Why should humans relinquish control entirely? Even in the case of human agents, complete independence is rarely accepted. Agents are typically subject to duties of disclosure and reporting. For example, *Article 924 of the Civil Code of the People's Republic of China* stipulates that an agent shall report the handling of entrusted affairs in accordance with the principal's requirements. During the course of agency activities, important information relevant to the entrusted matter must be communicated to the principal so that the latter remains informed of the progress of the undertaking.

Second, the criteria of "dangerous", "undesirable", and "beyond human competence" are themselves defined according to human perceptions, abilities, and preferences, rendering the scope of application highly flexible. Under such standards, virtually any work environment could become a justification for developing autonomous intelligent systems. Examples include autonomous vehicles, autonomous weapons, and algorithmic decision-making systems for stock trading. Autonomous weapons may appear to satisfy the criterion of danger, but what about autonomous driving systems or intelligent stock-trading systems? Admittedly, the value of technological artifacts is often assessed according to anthropocentric standards. Yet why should the demand for autonomous intelligent systems be grounded solely in such standards? Because these systems simulate and amplify a wide range of human capacities, they are potentially capable of undertaking any task that humans regard as dangerous, undesirable, or difficult. In this sense, they seem capable of satisfying human demands to an unprecedented degree. However, this aspiration is itself highly

controversial. Given that artificial agents emulate and extend human capabilities, Luciano Floridi characterizes artificial-agent technologies as part of the "Fourth Revolution" [11], while ecocentric thinkers regard them as yet another manifestation of humanity's enduring obsession with expanding technological control over the natural world.

## 4. Critiquing ethical safety issues within the framework of "technical problems"

### 4.1. Deconstructing the "technical problem": ethical safety

In professional technological practice, engineers are generally assumed to possess specialized expertise and a high degree of control over the artifacts they create. In the case of autonomous intelligent systems, however, professionals themselves remain preoccupied with the question of how to control the behavior of these systems. This reality challenges a fundamental assumption: that technical experts maintain a high level of control over their technological creations. Nick Bostrom introduced the concept of the "principal–agent problem" specific to artificial intelligence technologies [12], highlighting the fact that even experts such as engineers may be vulnerable when confronting advanced AI systems. Likewise, Paula Boddington argues that: "The distinctiveness of AI ethics lies in the fact that it extends beyond the scope of ordinary professional ethics. In terms of knowledge, understanding, and control, engineering communities are vulnerable in the face of autonomous intelligent systems; these systems are not fully controllable" [13]. If this is the case, then the ethical safety issues posed by autonomous intelligent systems can no longer be regarded simply as "problems in technology". This fact also serves as a reminder to end users and the broader public that autonomous intelligent systems urgently require public participation and social oversight.

Moreover, autonomous intelligent systems have introduced unprecedented technological phenomena: their operation is characterized by autonomy and opacity. The problems arising from such autonomy and opacity cannot be reduced to ordinary technical safety concerns. Rather, they provoke profound philosophical and ethical reflection. At this point, the engineering practice of autonomous intelligent systems encounters some of the most fundamental theoretical questions imaginable—questions concerning autonomy, human nature, agency, and responsibility. These are not merely technical issues but problems laden with ethical significance. Consequently, they cannot be adequately defined through the conventional framework of safety problems, even when modified by the adjective "ethical".

Indeed, describing these challenges as "ethical safety problems" risks reducing ethical concerns to matters of safety. Yet ethical issues and safety issues belong to different categories of inquiry. As Aimee van Wynsberghe and Scott Robbins observe: "Concepts such as values, rights, freedom, good and evil, and right and wrong lie at the heart of ethical inquiry and form the basis of competing conceptions of the good life. One may regard the values of safety and security as foundational to a good life; however, morality cannot be reduced to a problem of safety" [14].

### 4.2. Introducing "problems of technology": the problem of control

Once the presumed legitimacy of autonomous intelligent systems is called into question, the difficulties they present can no longer be understood merely as problems within the technology. Rather, they may be regarded

as problems of the technology[4] itself. The central problem of autonomous intelligent systems stems precisely from the pursuit of machine autonomy.

What, then, does machine autonomy mean? According to Vincent C. Müller, a system is autonomous to the extent that it operates independently of human control [15]. To distinguish the autonomy of autonomous intelligent systems from the mere automaticity of traditional machines, autonomy in this context may be defined more precisely as the capacity of an intelligent machine to act independently in the absence of human control, particularly with respect to decision-making and choice.

Paula Boddington has noted that the autonomy of autonomous intelligent systems gives rise to a distinctive control problem: "Certain degrees of machine autonomy[5] (or, more precisely, automaticity) may be compatible with control. For example, an autonomous missile can be programmed to target with greater precision than a bullet, yet once a bullet is fired, it is no longer under our control. But what happens when autonomy involves decisions and actions? The distinctive feature of AI often lies in its autonomy in judgment, decision-making, and action—concepts that are central to moral interpretation at a very deep level" [14].

In the case of ordinary technological artifacts, control problems typically arise because technology shapes the world in unpredictable or incomprehensible ways. With autonomous intelligent systems, however, a new and distinctive form of technological control problem emerges: how to control an actor that increasingly resembles a human being. First, the control problem of autonomous intelligent systems manifests itself as a potential conflict between machine-like human agency and uniquely human agency. Contemporary semi-autonomous and fully automated intelligent systems are already engaged in decision-making and choice. In this sense, they have become practical agents. Functionally, they stand alongside human decision-makers as counterparts. Yet this very functional equivalence constitutes an ethical issue that demands careful scrutiny. How can we ensure that such capabilities serve as effective extensions or supplements to human agency rather than replacing or undermining it?

Second, the issue extends beyond risk management. It also raises profound questions concerning ethical conceptual innovation. Faced with a human-like artificial actor, what kind of stance should human beings adopt toward its control? Should the relationship be one of master and servant? One of equal trust? Or should an entirely new ethical orientation be developed? These questions themselves constitute major ethical challenges. They have the potential to reshape many of the conceptual foundations of normative ethical theory. The problem of control therefore exceeds the explanatory scope of technological rationality. It possesses a deeply ethical character and demands that autonomous intelligent systems be approached with considerable caution and philosophical seriousness.

[4] The term "problems of technology" is employed in opposition to "problems in technology" discussed above. See Footnote 2 for a detailed explanation of this distinction.

[5] In English, both "zi zhu (自主)" and "zi dong (自动)" are often rendered as autonomy. However, the example provided by Paula Boddington clearly suggests an important distinction. Automaticity may refer to forms of behavior that are non-human in nature, whereas autonomy appears to involve modes of action characteristic of human agents, particularly the capacities for judgment, decision-making, and choice.

# 5. Critiquing the radical technocratic strategy of moral machines

## 5.1. Addressing the control problem through caution and restraint

The challenge posed by autonomous intelligent systems is not merely an ethical safety threat in the conventional sense. Rather, it is a control problem with profound ethical implications arising from the autonomy of these systems. Consequently, such a problem should be approached with caution and restraint.

One possible response is to establish dedicated regulatory institutions to oversee the development of autonomous intelligent systems from the outside. Such institutions could regulate the development of autonomous systems from a broader and longer-term perspective and, where necessary, prohibit or restrict certain forms of development. For example, the European Parliament's recommendations on civil law rules for robotics state that research may require regulation or even prohibition when autonomous robots pose threats to human safety. The recommendations further propose the future establishment of a European agency for robotics and artificial intelligence to identify potential risks [16].

A second approach is to adopt restrictive safety strategies that limit the autonomy of autonomous intelligent systems. The concept of "AI Safety Engineering" proposed by Roman V. Yampolskiy [8] exemplifies such a restrictive technological strategy. Several concrete proposals have already been advanced. Eric Drexler suggests confining an autonomous agent within "sealed hardware", allowing researchers to safely study and utilize its outputs while preventing it from causing harm to human beings [17]. Nick Bostrom proposes the concept of an Oracle AI (OAI), an artificial intelligence system that answers questions but does not make decisions or take actions [12]. David Chalmers advocates a "leakproof" approach [18], under which AI systems are initially restricted to simulated virtual environments until their behavioral tendencies can be fully understood under controlled conditions. Similarly, Yampolskiy proposes the "Artificial Intelligence Confinement Problem" (AICP) [19], whereby an AI agent is restricted to a bounded environment and prevented from communicating with the outside world unless such communication is explicitly authorized by the containment system.

## 5.2. Moral machines as a radical and proactive technological strategy

The moral-machine project, by contrast, represents a radical and proactive technological strategy. In engineering practice, safety defects are typically addressed through restrictive or preventive safety technologies. Classical safety technologies generally function by limiting either the circumstances under which a technology operates or the manner in which it operates. Safety-by-design measures exemplify restrictions on modes of operation. Elevator doors, for instance, are equipped with sensors that prevent them from closing on people, while lawnmowers contain protective mechanisms that shield users from rotating blades. Restrictions may also be imposed through environmental controls, such as the safety barriers used in logistics facilities to separate pedestrians from vehicles.

Compared with these traditional restrictive safety strategies, moral-machine technology is considerably more proactive. The objective of the moral-machine project is to develop entities capable of moral decision-making and thereby enhance the capacity of autonomous intelligent systems to navigate ethical dilemmas. As a positive safety strategy, it seeks to preserve the autonomy of autonomous intelligent systems to the greatest possible extent. In contrast, approaches such as the "leakproof" containment model, Artificial Intelligence Confinement Problem (AICP), and Oracle AI (OAI) are more consistent with restrictive safety strategies. These proposals limit—or even prohibit—certain modes of use and operational contexts, thereby reducing the autonomy of intelligent systems.

There are also examples of positive safety strategies that resemble the moral-machine approach. Modern trains, for instance, may incorporate fail-safe systems designed to respond to emergencies such as driver incapacitation. Such systems constitute positive safety measures because they activate intelligent decision-making mechanisms under specific circumstances. Yet although both approaches expand the operational scope of autonomous systems, the moral-machine project is substantially more radical. Unlike a train's fail-safe system, a moral machine introduces an entirely new dimension of autonomy: the capacity for moral decision-making.

Even if one sets aside the argument that autonomous intelligent systems ought to be approached with caution and restraint, the choice of moral machines as a technological solution remains questionable from within a technocratic framework itself. By departing from established paradigms of safety engineering, the moral-machine project adopts a strategy that is itself sufficiently radical to warrant careful scrutiny.

### 5.3. Critiquing the instrumentalist presupposition of the moral-machine project

As a form of subjective rationality, technological rationality entails a degree of agency, primarily reflected in the active selection and balancing of means and ends. This agency presupposes that human beings exercise complete control over technological means and that technology functions merely as an instrument employed by humans. The moral-machine project is grounded precisely in this instrumentalist conception of technology. As Steve Torrance observes: "It is an anthropocentric position. No matter how intelligent autonomous systems become or how closely they resemble human beings, they remain nothing more than tools used by humans. Machine ethics is essentially a form of safety-systems engineering, or an extension thereof" [2].

Autonomous intelligent systems, however, challenge the assumptions of technological instrumentalism. Such systems are not simple tools; rather, they increasingly resemble human-like agents. John P. Sullins argues that: "Some contemporary robots have more in common with guide dogs than with simple tools such as hammers. The question is whether robots should properly be regarded as just another kind of tool, or whether they more closely resemble the kind of technology exemplified by guide dogs. Even at the current stage of robotics, this is not an easy question to answer" [2].

The autonomy of autonomous intelligent systems gives rise to a control problem with significant ethical implications. Nevertheless, practical machine ethics continues to adhere to an instrumentalist conception of technology and confines the goal of moral-machine development to the domain of instrumental utility. Yet autonomous intelligent systems are not merely tools, and moral machines are even less so. Unlike traditional safety technologies, moral machines are far more radical and proactive, possessing a substantial degree of autonomy. By endowing already problematic autonomous intelligent systems with moral agency, practical machine ethics does not resolve the control problem; rather, it intensifies it.

## 6. The rebellion of moral machines against technological rationality

By endowing autonomous intelligent systems with ethical agency, the moral-machine project deepens the control problem inherent in such systems. In doing so, it ultimately departs from the fundamental objective of technological rationality: the realization of human-centered control and transformation of the world in accordance with human values.

### 6.1. The loss of control over moral machines

Moral machines may ultimately escape human control and become autonomous forces that stand opposed to their creators. In other words, we may lose technological control over moral machines themselves. In their

article *Moral Control and Ownership of Artificial Intelligence Systems*, Raul Gonzalez Fabre et al. argue that the complexity of artificial intelligence systems often means that decisions emerge from interactions among multiple systems, designers, and stakeholders [20]. Under such circumstances, it becomes difficult to determine where moral agency originates or whose moral agency is being expressed. As a result, human beings may lose effective moral control over AI systems. The introduction of artificial moral agents into human ethical life—and the replacement or restriction of one or more aspects of human moral activity by such agents—may ultimately produce consequences that are opaque, unpredictable, and beyond human control. The "behavior" of artificial agents may effectively escape the domain traditionally governed by human morality.

Like all technologies, moral machines are susceptible to failure and loss of control. However, their risks differ from those associated with ordinary technologies because they belong to the category of high-risk technologies. Nuclear weapons, for example, are technologically mature and can be controlled with considerable precision. Yet this does not guarantee that their operation will always remain within human expectations. During international conflicts or in situations involving human error and institutional failure, nuclear technologies may become sources of catastrophic risk. As Miles Brundage notes: "People can design moral robots according to their own preferences.… In a world where unethical individuals exist and can gain access to advanced technologies, technological solutions proposed by machine ethics may ultimately be of little significance" [21].

## 6.2. The threat of moral outsourcing to human-centered status

As a technological project, the moral machine transforms human ethical judgment into a technical and instrumental process. Compared with technologies such as elevator sensors, leakproof containment mechanisms, or fail-safe systems embedded in trains, moral machines represent a fundamentally different category of technology. Miles Brundage argues: "Moral-machine technology entails substantial risks. The challenge of advanced AI should be addressed from a broader range of perspectives, including hardware constraints, regulation of advanced AI systems, and the cultivation and reinforcement of human values, rather than assuming that our moral intuitions can soon be systematically formalized and programmed" [23].

The development of moral machines may introduce moral agents other than human beings into ethical life, thereby threatening the privileged status of human moral agency. John Danaher argues that: "The moral agency of artificial intelligence and robots may suppress human moral agency, gradually reducing human beings to moral patients. If moral agency is regarded as an important civilizational value, then artificial moral agency may pose deeper civilizational risks" [22].

Although practical machine ethics is grounded in an anthropocentric form of technological rationality, its proposed solution may ultimately contribute to human decentering. Artificial intelligence has consistently compelled human beings to reconsider their own nature and status. Like all technological artifacts, AI represents a form of outsourcing human capabilities. Yet unlike previous technologies, AI externalizes capacities that are central to human identity—namely, the capacities for judgment, choice, and decision-making. In doing so, it challenges the Cartesian conception of the human being. As an intelligent agent, artificial intelligence has already altered established frameworks for understanding human uniqueness, including the notion that human beings are uniquely rational animals and possess distinctive forms of intelligence. Practical machine ethics goes even further by proposing the outsourcing of ethical decision-making itself, thereby expanding the scope of capability outsourcing into one of the most fundamental domains of human existence. This raises an important question: does such outsourcing have limits? The issue recalls the philosophical thought experiment of the experience machine. If human beings ultimately delegate even the most important ethical decisions to machines, do we wish to become passive experiencers whose

moral lives are mediated by external systems? Such a possibility opens a series of profound philosophical questions. Different interpretations of the moral functions assigned to machines imply different conceptions of human nature and different visions of the ethical world. The outsourcing of moral agency therefore forces us to reconsider not only the future of technology but also the meaning of moral responsibility, human autonomy, and the kind of ethical community in which we wish to live.

## 7. Conclusion

"Rationality should not be manifested solely in regulating the relationship between ends and the means used to achieve them. It should also be reflected in the proper understanding and evaluation of those ends, as well as in the anticipation and assessment of the consequences of actions directed toward them. In short, rationality should be understood as humanity's capacity to choose and regulate its own conduct" [23]. Practical machine ethics fails to adequately recognize the distinctive nature of autonomous intelligent systems—their status as human-like agents. It remains firmly rooted in the framework of technological rationality, treating such systems as ordinary tools and attempting to control them through a new technological means: the moral machine. What this approach overlooks is that the ethical decision-making dilemmas faced by autonomous intelligent systems stem precisely from their autonomy and constitute control problems with profound ethical significance, rather than ordinary safety problems. Moreover, moral-machine technology is not a conventional safety technology. Unlike traditional restrictive safety technologies, which operate by limiting or constraining technological behavior, moral-machine technology is a proactive and radical safety strategy. Paradoxically, this strategy intensifies rather than resolves the control problem. At the same time, as a technological instrument, machine morality exceeds the conventional means–end framework and introduces significant technological risks. It also generates new ethical risks, including the outsourcing of human moral capacities and the corresponding transformation of conceptions of human nature and the ethical order. In short, practical machine ethics lacks sufficient critical reflection on autonomous intelligent systems themselves. It reduces the challenge of controlling autonomous intelligent systems to the narrower question of how to ensure that such systems behave ethically. Yet the control problem is a broader and more profound issue in AI ethics. It concerns not only the status of autonomous intelligent systems as agents but also the challenge they pose to human subjectivity and agency.

Autonomous intelligent machines have already begun to affect the very core of the human lifeworld—the moral domain. The question, however, is how we ought to respond to this transformation. Michael Anderson and Susan Leigh Anderson argue that it would be highly imprudent and unwise to develop autonomous intelligent machines that lack moral sensibility [2]. By the same token, this paper contends that the development of moral machines is itself equally imprudent and unwise. Before embarking upon such an ambitious engineering project, it is necessary to engage in a deeper examination of two fundamental questions: how the tension between human agency and the human-like agency of autonomous intelligent systems should be managed, and, more fundamentally, what kind of beings human beings ultimately aspire to become.

## Funding project

Youth Project of the Chongqing Social Science Planning Program, "Polanyi's Concept of Intelligence and Its Implications for Artificial General Intelligence Research" (2019QNZX48); the Doctoral Cultivation Project of the Central Universities' Fundamental Research Funds, "A Study of the Crisis of Trust from the Perspective of Modernity" (SWU1509503); the General Teaching Reform Project of Southwest University, "Methodological

Exploration of Incorporating Engineering Ethics into Engineering Education Systems" (2024JY047); and the Southwest University Innovative Research 2035 Pilot Plan Project (SWUPilotPlan018)

## References

[1] Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.
[2] Horkheimer, M. (1947). *Eclipse of reason*. Oxford University Press.
[3] Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A. F., & Yampolskiy, R. V. (2017). *Towards moral autonomous systems*. arXiv. https://doi.org/10.48550/arXiv.1703.04741
[4] Pereira, L. M., & Lopes, A. B. (Eds.). (2020). *Machine ethics: From machine morals to the machinery of morality*. Springer International Publishing.
[5] Wang, D. Z., & Guan, S. X. (2004). Philosophy of technology, technological practice, and technological rationality [in Chinese]. *Philosophical Research*, *11*, 56–63.
[6] Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–25.
[7] Allen, C., & Wallach, W. (2010). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
[8] Yampolskiy, R. V. (2018). *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*. In V. C. Müller (Ed.), Philosophy and theory of artificial intelligence (pp. 388–395). Springer International Publishing.
[9] Feenberg, A. (2005). *Questioning technology (H. L. Q. Han & G. F. Cao, Trans.) [Chinese translation]*. Peking University Press. (Original work published 1999)
[10] Whitby, B. (1996). *Reflections on artificial intelligence: The legal, moral, and social dimensions*. Intellect; Cambridge University Press.
[11] Floridi, L. (2008). Artificial intelligence's new frontier: Artificial companions and the fourth revolution. *Metaphilosophy, 39*(4–5), 651–655.
[12] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
[13] Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer International Publishing.
[14] van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(2), 725–735.
[15] Müller, V. C. (2012). Introduction: Philosophy and theory of artificial intelligence. *Minds and Machines, 22*(2), 67–69.
[16] European Parliament, Directorate General for Internal Policies. (2016). *European civil law rules in robotics.* European Union.
[17] Drexler, K. E. (1986). *Engines of creation: The coming era of nanotechnology*. Anchor Press.
[18] Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9–10), 7–65.
[19] Yampolskiy, R. V. (2012). Leakproofing the singularity: Artificial intelligence confinement problem. *Journal of Consciousness Studies, 19*(1–2), 194–214.
[20] Fabre, R. G., Ibáñez, J. C., & Escobar, P. T. (2021). Moral control and ownership in AI systems. *AI & Society, 36*(1), 289–303.
[21] Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence, 26*(3), 355–372.
[22] Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society, 34*(1), 129–138.
[23] Gao, L. H. (1993). An exploration of the philosophy of technology and the problem of technological rationality [in Chinese]. *Philosophical Research*, *2*, 78–84.