Available Online: 14 September 2026 DOI: 10.54254/2753-7080/2026.36869

# Regulative but not distributive: a dual-layer model of artificial moral responsibility

*Xiaoke Xia*

Shi Cheng College, Hunan Normal University, Changsha, China

minicle@126.com

**Abstract.** As autonomous AI systems increasingly cause real-world harm, traditional responsibility frameworks—anchored in free will and the capacity to suffer—have proved inadequate, resulting in what scholars call a "responsibility gap". This study explores the conceptual viability of artificial moral responsibility and delineates its definitional boundaries. Drawing on responsibility theories, functionalist agent frameworks, and punishment theory, alongside psychological findings of retributive intuitions and sociological insights into collective conscience, this paper differentiates regulative responsibility from distributive responsibility and systematically evaluates the extent to which artificial intelligence meets the constitutive criteria of regulative moral responsibility. The analysis shows that AI can legitimately bear corrective responsibility through consequentialist correction mechanisms, yet cannot satisfy victims' retributive demand for desert, which requires a morally responsive human community. The paper concludes by proposing a "dual-layer responsibility model", in which AI assumes corrective responsibility while humans retain residual symbolic responsibility—jointly addressing both the responsibility gap and the retribution gap without fabricating AI minds or abandoning victims' demands for justice.

**Keywords:** artificial moral responsibility, responsibility gap, retribution gap, dual-layer responsibility model

## 1. Introduction

The rapid deployment of autonomous AI systems in high-stakes domains—autonomous vehicles, algorithmic decision-making, and even military applications—has made the question of moral responsibility for AI-caused harm increasingly urgent. Existing scholarship has largely converged on a troubling diagnosis: as systems become more autonomous, neither their designers nor their operators can be fairly held responsible for outcomes they could not predict or control, producing what Matthias terms a "responsibility gap" [1]. Yet most existing responses either anthropomorphize AI by attributing to it dubious mental states, or pessimistically conclude that responsible AI deployment is simply impossible. What remains underexplored is a middle path: a responsibility framework that neither fictionalizes AI minds nor abandons the demand for accountability. This paper addresses that gap by asking whether artificial moral responsibility is possible, and if so, what form it must take to remain philosophically defensible. Specifically, it investigates whether responsibility can be reconceived in functional, process-based terms rather than property-based terms, and whether such a reconception allows AI to bear one kind of responsibility while leaving another kind irreducibly human.

Through conceptual analysis drawing on responsibility theory, functionalist agency theory, and punishment theory, this paper develops a framework intended to inform both AI governance design and future philosophical debate on machine accountability.

## 2. Distinguishing between concepts of responsibility

### 2.1. The distinction between distributive responsibility and regulative responsibility

Before asking whether AI can bear moral responsibility, it must first clarify what kind of responsibility is at issue. This paper argues that moral responsibility is not a unitary concept but consists of two distinct forms with different normative functions.

Current discussions on AI accountability can be broadly categorized into two schools of thought: Distributive responsibility and regulative responsibility. Distributive responsibility focuses on the blame, punishment, or praise that an agent "deserves" based on past actions. Grounded in retributive justice, it requires the agent to possess free will, moral judgment, and the capacity to suffer, since punishment here is a fitting response to past wrongdoing rather than a means to some future end. It also serves important symbolic and socio-emotional functions. Robert Sparrow has argued that autonomous weapons systems ("killer robots") create an insurmountable responsibility gap [2], since machines lack free will and the capacity to feel pain. Danaher's related concept of the "retributive void" similarly highlights the tension between society's desire for retributive punishment and the lack of an agent capable of bearing it [3].

Regulative responsibility, however, is different; rather than asking who "deserves" what, it focuses on regulating future behavior to prevent recurrence. Daniel W. Tigard explicitly opposes the notion of a so-called "technological responsibility gap": moral responsibility, on his view, is a dynamic process of social interaction, so AI can still be "held accountable" through rewards and punishments such as restricting permissions or retraining [4]. This form of responsibility is grounded in the corrective logic of consequentialism; it does not require the responsible entity to possess free will or the capacity to suffer, but only that it be capable of perceiving external signals and alter its behavioral patterns accordingly. Corrective measures such as retraining or restricting decision-making authority are functionally similar to "punishment", but their legitimacy lies in reducing future risk rather than in retribution. Responsibility here is a regulatory mechanism oriented toward behavioral correction, into which AI can be functionally integrated as an appropriate subject of regulative responsibility. However, this can only be a limited form of responsibility and cannot fulfill the full functions of retributive justice.

### 2.2. The reason why artificial intelligence is suitable for regulative responsibility

In what sense, then, can AI systems serve as appropriate bearers of regulative responsibility? AI is structurally capable of meeting the formal conditions required by regulative responsibility; in fact, in certain dimensions, it is even better suited than humans.

Regulative responsibility can be separated from mental predicates such as free will and the capacity to suffer, as its criteria are essentially functional: the bearer of responsibility need only possess traceability, intervenability, role embedding, and feedback sensitivity to support "regulation to prevent recurrence". This study examines AI's performance across these four dimensions below.

First is traceability. An AI's decision-making path is a deterministic mapping of parameters, training data, and algorithmic rules to a given input; its causal chain can be fully reconstructed through logs and explainability tools.

Second is intervenability. An AI's behavioral patterns can be directly altered through retraining or parameter fine-tuning. This is operational, unlike interventions on human character, which rely on education and remain merely advisory.

Third is role-embeddedness. The core intuition behind Hart's "role-based responsibility" is that responsibility is tied to the position an agent occupies within a larger system rather than to mental states [5]; autonomous driving modules occupy comparable structural positions and thus possess the prerequisites to bear regulative responsibility.

Fourth is feedback sensitivity. AI's response to feedback signals (gradient updates, reward adjustments) is itself central to its operational mechanism; it lacks the psychological resistance common in humans, and feedback sensitivity is virtually inherent to the very nature of AI.

Taken together, these four dimensions show that AI not only meets the conditions for regulative responsibility, but in many respects, demonstrates a higher degree of alignment than humans. This suggests that regulative responsibility may have been designed from the outset for functional agents rather than mental agents, and AI is merely a case that makes this point explicit.

## 3. An argument regarding the possibility of artificial moral responsibility

### 3.1. From the property view to the process view

To answer the question "Can artificial intelligence bear regulative responsibility?", this paper must first address a more fundamental question: What qualifications must the bearer of responsibility satisfy? Traditional responsibility theories adopt a property view, according to which responsibility presupposes free will, intentionality, or the capacity to suffer.

This view faces two dilemmas when applied to AI. The first is an empirical dilemma: it cannot determine whether AI possesses intentionality or phenomenal consciousness, since its "human-like" qualities may well be merely the product of statistical fitting. The second is a deeper theoretical dilemma: even assuming AI lacks mental states, the property view cannot explain the fact that real harm caused by AI still exists. Since self-learning machines render their future behavior unpredictable to their creators, who cannot then be fairly held morally accountable, it is left to acknowledge the "responsibility gap" that traditional concepts of responsibility can no longer bridge [1]. The attribute view is the theoretical root of this gap: once it links responsibility too tightly to mind and denies AI has one, it can only shift responsibility back to humans.

Faced with this dilemma, a shift toward the "process view" becomes both reasonable and necessary: rather than prioritizing the question of "who is what kind of being", it should instead ask who or what occupies the nodes within the causal chain where correction can be appropriately applied. Responsibility is thus not a static attribute but a functional role activated within a system's operation. This process view is consistent with Dennett's functional account of responsibility and Floridi and Sanders' conception of artificial agency [6, 7].

This allows us to reposition the two stances: the mental premise required by the attribute-based view is not universal, but merely a condition for distributive responsibility specifically—only when asking "who deserves punishment" do free will and the capacity to suffer become irreplaceable. Regulative responsibility has never been about "deserving", but rather about "correction" and "preventing recurrence"; therefore, it requires only the functional, observable, and intervenable conditions for a bearer of responsibility as characterized by the process-oriented view—conditions that Part Two already demonstrated AI can satisfy.

However, the scope opened up by the process-oriented view is not boundless: an air conditioner thermostat also reacts to environmental changes, yet it is not regarded as a responsible agent. Causal links alone are thus insufficient.

### 3.2. AI meets the conditions for a vehicle of responsibility

This paper then shifts the focus to the more basic theoretical research centered on the process-oriented framework: identifying the internal system attributes that make an entity truly bear moral responsibility. This paper proposes three progressive criteria. The first is boundary locatability. A bearer of responsibility must be a functional unit with clearly defined boundaries and explicit input and output interfaces. The thermostat satisfies only this criterion. The second is internal complexity and adaptability—specifically, interactivity, autonomy, and adaptability, as proposed by Floridi and Sanders [7]. Thermostats fail this criterion, whereas AI systems with learning capabilities satisfy it. The third criterion is the practical effectiveness of accountability. Corrective measures imposed on this unit must actually alter its future behavior patterns, rather than falling flat or being circumvented—this excludes automated scripts with fixed rules.

These three criteria are not tailored specifically for AI; the legal person liability system has long demonstrated their general applicability: French points out that corporations qualify as moral agents by virtue of their internal decision-making structures and rules [8]—corporations have no unified mind, yet are recognized as appropriate bearers of liability, confirming that the bearer of liability need not be a natural person.

However, "being capable of serving as a vehicle for liability" and "what nature of corrective measures should be imposed" remain two distinct issues; the next section turns to consequentialist theory of punishment to ask which corrective logic applies to AI.

### 3.3. The application of consequentialism punishment

As a bearer of responsibility, what nature of corrective measures should AI be subject to? Punishment theory has historically been divided into two major camps: retributivism asserts that the legitimacy of punishment lies in the idea that punishment fits the crime, and its line of argument is retrospective. The consequentialist theory of punishment, on the other hand, maintains that the justification for punishment lies entirely in its future effects (deterrence, rehabilitation, and social defense), tracing back to Bentham's utilitarian view and adopting a forward-looking line of argument.

This distinction corresponds precisely to the attributivist/process-oriented distinction in Section 3.1: retributivism's justificatory foundation rests on the mental premises required by the attributivist view, whereas consequentialism requires only the functional conditions characterized by the process-oriented view. Since AI cannot satisfy the premises of the attributivist view, it cannot be suited to punishment in the retributivist sense, but it is precisely suited to correction in the consequentialist sense.

However, consequentialism has long been criticised for permitting the punishment of the innocent [9]: if justification depends entirely on future effects, punishing an innocent party is logically justifiable too, contradicting the intuition that punishment must be predicated on fault. This limitation is magnified when the subject of correction is an AI: lacking an intrinsic perspective, correction in its case is closer to "repair" than "punishment", and however effective, cannot satisfy victims' demand that someone be truly held accountable.

## 4. The two-tiered responsibility model

### 4.1. The demand for retribution cannot be eliminated

However, the fact that AI can withstand consequentialist corrective measures does not mean that retributive demands can be eliminated.

Psychological evidence suggests that people's intuitions about punishment are primarily driven by considerations of desert rather than by expected future consequences. Carlsmith et al. found that the severity of punishment closely tracks the seriousness of the offense, while varying little with the likelihood of future deterrence, indicating that retributive intuitions remain central to ordinary moral judgment [10]. This finding is reinforced by sociological and philosophical accounts. Durkheim argues that punishment serves not only to deter wrongdoing but also to reaffirm the moral order of a community, a function that presupposes the participation of members who share that community's normative commitments [11]. Similarly, Sparrow maintains that punishment is meaningful only when directed at an entity capable of participating in the moral relationship between punisher and punished [2]. Since AI is not a member of the human moral community, corrective measures alone cannot satisfy the demand for retributive justice.

It is thus evident that the irreducibility of the demand for retribution is supported at both the individual psychological level and the level of social theory. No matter how effective consequentialist corrective measures may be, they can only address the issue of "preventing future occurrences"; they cannot address the independent demand that "past wrongs must be acknowledged by an entity capable of understanding condemnation".

## 4.2. Two levels of responsibility: AI's corrective responsibility and humans' symbolic responsibility

Section 4.1 has demonstrated that the consequentialism corrective measures AI can undertake address the responsibility gap but fail to address the retributive gap, so a single level of responsibility cannot fill both. This section proposes a "two-tier responsibility model": AI bears corrective responsibility, while humans bear symbolic responsibility.

Corrective responsibility is the institutional implementation of the chain of "traceability—intervenability—verifiable behavioral change following correction" argued in Parts II and III. Its specific forms include: mandatory retraining or fine-tuning for specific failure patterns; taking the system offline or deploying it with restricted permissions; obligations to retain algorithmic audits and decision logs; and rule patches and version rollbacks for similar risks. This level is forward-looking and technical; accountability lies with the system itself, not with an individual imagined as a "culprit".

Symbolic responsibility, on the other hand, addresses two questions that corrective responsibility cannot resolve: who bears the responsibility, and what exactly does that responsibility entail. The bearer need not be a specific individual for whom a causal chain can be definitively established—this ambiguity is itself the source of the responsibility gap. The bearer should instead be the system's "personified representative" within the social institutional framework—an institution's head, a regulatory signatory, or the corporate entity itself—whose qualification stems from positional role rather than irrefutable causal fault.

In terms of content, symbolic responsibility encompasses at least three specific forms. The first is public moral acknowledgment: formally acknowledging the occurrence and severity of the harm through gestures that can be interpreted as remorse or apology. Second is the bearing of tangible costs: resignation, reputational damage, financial compensation, or legal liability, satisfying the demand that "someone has indeed paid a price" (Section 4.1). Third is ritualistic procedures—public hearings, statements of accountability, commemorative arrangements—echoing Durkheim's insight that punishment affirms shared moral order.

The two forms of responsibility complement each other. Corrective responsibility reduces the risk of future harm, while symbolic responsibility acknowledges past wrongdoing and responds to victims' demand for accountability. Together, they address both the responsibility gap and the retribution gap without requiring AI to possess human-like mental capacities.

## 5. Conclusion

This paper has argued that artificial moral responsibility is possible, but only in a regulative rather than a distributive sense. Rather than asking whether AI possesses free will or the capacity to suffer—prerequisites that have long anchored distributive responsibility—this paper shifted the question from a property view to a process view, asking instead whether AI occupies an appropriate functional position within a corrective process. On this basis, the paper showed that AI satisfies the conditions for serving as a responsibility-bearer: its behavior is traceable, modifiable, embedded in a functional role, and responsive to feedback in ways that consequentialist corrective measures can exploit.

Yet this corrective capacity, however effective, cannot do everything that responsibility is asked to do. Drawing on psychological evidence showing that retributive intuitions persist independently of deterrent outcomes, and on sociological accounts of punishment as a community's reaffirmation of shared moral order, the paper showed that victims' demand for desert cannot be satisfied by a system with no inner life capable of being meaningfully condemned. This is the retribution gap that purely technical correction leaves untouched.

The dual-layer responsibility model proposed here—AI bearing corrective responsibility, humans retaining residual symbolic responsibility—offers a way of honoring both halves of this problem without overstating what AI is or understating what victims deserve. As autonomous systems take on greater decision-making authority, this framework suggests that accountability structures should be designed deliberately around this division, rather than collapsing it into either uncritical optimism about machine accountability or a reflexive insistence that only humans can ever be answerable for AI-caused harm.

## References

[1] Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.

[2] Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*, 62–77.

[3] Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*, 299–309.

[4] Tigard, D. W. (2021). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics, 30*, 435–447.

[5] Hart, H. L. A. (1968). *Punishment and responsibility: Essays in the philosophy of law*. Clarendon Press. Oxford.

[6] Dennett, D. C. (2015). Mechanism and responsibility. In T. Honderich (Ed.), *Essays on freedom of action (Routledge Revivals)*. Routledge. Boston.

[7] Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379.

[8] French, P. A. (2022). The corporation as a moral person. In P. Jones (Ed.), *Group rights*. Routledge. London.

[9] McCloskey, H. J. (1965). A non-utilitarian approach to punishment. *Inquiry, 8*, 249–263.

[10] Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*, 284–299.

[11] Durkheim, E. (2023). The division of labor in society. In *Social theory rewired*. Routledge. New York.