Cheng, P., and Zh. Zhang*. 2025. "The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 123–151.

Peng Cheng and Zhihui Zhang*

The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence**

Submitted: Sept. 13, 2025. **Reviewed:** Sept. 30, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This article addresses the "responsibility gap" arising from the integration of embodied artificial intelligence into social interactions. Rejecting functionalist models that equate AI agency with moral personhood, we adopt a phenomenological perspective, reframing responsibility as a relationally "manifested" phenomenon. We propose a Three-Stage Model of Embodied Responsibility Emergence (Perceptual Presentation — Situational Embeddedness — Ethical Appeal) and philosophically reconstruct the four dimensions of RRI (Anticipation, Reflexivity, Inclusion, Responsiveness) as a "Structure of Responsibility Manifestation." The paper warns against the ethical risks of anthropomorphism, clarifies AI's inherent lack of empathy and moral agency, and argues that responsibility must ultimately reside with humans and be institutionalized through governance mechanisms. Operable pathways and normative boundaries for such governance are outlined.

Keywords: Embodied Artificial Intelligence, Responsibility Attribution, Ethical Governance, Phenomenology of Responsibility, Responsible Research and Innovation.

DOI: 10.17323/2587-8719-2025-4-123-151.

Today's embodied artificial intelligence (Embodied AI) systems, such as robots, are deeply integrated into social interactions. From companion assistants to autonomous vehicles, they are increasingly acting autonomously and influencing the human world. However, such AI does not possess human-like subjective consciousness or moral intent. Since responsibility has traditionally been attributed to moral agents with autonomous will, how an AI without subjectivity can bear the moral responsibility for the consequences of its actions has become an ethical dilemma (Coeckelbergh, 2020). For

*Peng Cheng, PhD in Philosophy; Associate Researcher at the China Research Institute for Science Popularization (Beijing, China), cheng80416519@126.com, ORCID: 0009–0002–8474–4711; Zhihui Zhang, PhD in Philosophy; Professor at the Institute for History of Natural Sciences, Chinese Academy of Sciences (Beijing, China), zhangzhh@ihns.ac.cn, ORCID: 0000–0003–1876–9312. Corresponding author: Zhihui Zhang.

example, when an autonomous vehicle is involved in an accident, should the responsibility be attributed to the machine itself, its manufacturer, the user, or regulatory entities? This "responsibility gap" resulting from the increasing autonomy of artificial intelligence has drawn significant academic attention (Matthias, 2004). In the field of AI ethics research, on one hand, some scholars attempt to apply models of moral agency to AI, arguing that as long as AI demonstrates human-like functionalities, it can be regarded as a responsible agent; on the other hand, others emphasize that AI lacks intrinsic subjectivity, and therefore responsibility must ultimately reside with humans. However, as embodied AI assumes more active interactive roles in society, this simplistic dichotomy has become inadequate for addressing complex real-world scenarios.

Current discussions often fall into the trap of functionalism, evaluating AI's ethical capabilities through the lens of behavioral equivalence to subjective consciousness. For instance, Turing Test-style logic suggests that if AI's external behavior is indistinguishable from humans, it may be granted corresponding moral status. This has led some researchers to propose a "Moral Turing Test," where humans distinguish between machine and human ethical judgments. Recent experiments have shown that moral persuasion responses generated by large language models are sometimes perceived as more virtuous than human responses, thus in a sense "passing" the Moral Turing Test (Georgia State University News Hub, 2024). This reveals the risks of relying solely on behavioral functionality to determine AI's moral identity: AI can "deceive" us by simulating rationality and compassion, while in fact lacking genuine moral understanding or emotions. This functionalist perspective tends to induce attribution errors: people are inclined to attribute consequences caused by AI to the AI itself, thereby obscuring the human agents who should actually bear responsibility. A study demonstrates that when interactions with robots fail, people often tend to blame and assign responsibility to the robots due to attributing human-like mental capacities to them (Kawai et al., 2023). Therefore, if responsibility is assigned based solely on AI's functional performance, it may lead to erroneous attribution of responsibility and ethical confusion.

Hence, the core of the issue resides in embodied AI's profound involvement in social actions despite its lack of subjective intentionality, which renders traditional responsibility attribution paradigms — predicated on agent intentionality and behavioral control — largely inadequate (Coeckelbergh, 2020). In response, this paper advocates transcending the limitations

of functionalist misinterpretations and attribution models, proposing a phenomenological reexamination of responsibility's generative mechanisms and governance logic.

The argument will unfold as follows: First, we delineate the functionalist model of moral agency and its consequent responsibility dilemmas, exposing its inherent logical deficiencies. Next, we introduce phenomenological philosophy to contend that ethics should not be reduced to functional attributes but must be understood as emergent within relational interactions. Building upon this foundation, we propose a "Three-Phase Model of Embodied Responsibility Generation," elucidating how responsibility progressively manifests through three interconnected processes: perceptual presentation, situational embeddedness, and ethical call. Subsequently, we explore the institutional embedding of this non-agential responsibility framework, sketching the contours of embodied AI ethical governance through a philosophical reconstruction of the four dimensions of Responsible Research and Innovation (RRI). Finally, we address potential critiques — including concerns about AI anthropomorphism and counterarguments regarding AI's incapacity for empathy — to clarify the boundaries of responsibility attribution. We emphasize that responsibility resides not within AI itself, but in humanity's response to the call of the Other.

Through this theoretical articulation, the paper seeks to establish a foundation for a "non-agential ethics" framework, offering new philosophical directions for the future governance of embodied artificial intelligence.

THE PROPOSAL AND CONNOTATION OF EMBODIED ARTIFICIAL INTELLIGENCE AND ITS ETHICAL DIFFERENCES FROM BRAIN-INSPIRED ARTIFICIAL INTELLIGENCE

The concept of Embodied Artificial Intelligence (EAI) emerged in the 1980s and 1990s, arising from reflections on and critiques of traditional Good Old-Fashioned AI (GOFAI), which relied on symbolic reasoning. Researchers such as Hubert Dreyfus, Rodney Brooks, and Rolf Pfeifer, among others, argued that truly intelligent artificial systems should not depend solely on abstract cognitive reasoning or algorithmic control. Instead, they must possess the capacity for closed-loop interaction encompassing "body-environment-perception-action"— that is, the ability to achieve understanding and adaptation with embodiment as a core premise.

From this perspective, embodied artificial intelligence is not merely a technical approach but also a philosophical rethinking of the nature of intelligence. Its core tenets include:

(1) emphasizing the body as the foundational medium of cognition, positing that intelligence stems from sensorimotor coordination rather than symbolic manipulation;
(2) stressing situational dependency, contending that intelligence is not a product of internal computational modules but rather a system's adaptive capacity within a dynamic environment;
(3) advocating that cognition is distributed, interactive, and embedded, rather than centrally localized within a processing unit.

In contrast, Brain-Inspired AI seeks to emulate the neural architecture of the human brain through neural network models — such as deep learning and cortical simulations — to replicate human perceptual and decision-making processes. Although both approaches share the goal of modeling "human intelligence," they differ significantly in their methodologies and philosophical assumptions: Brain-Inspired AI focuses on mapping the "information processing structure," whereas Embodied AI prioritizes the reconstruction of the "body-world interaction structure."

This divergence is particularly pronounced in discussions of ethical responsibility. Brain-inspired AI continues the cognitivist tradition, wherein its potential for responsibility is primarily conceived as stemming from its "agent-like internal structures" — such as the capacity for autonomous will, rational decision-making, and motivational judgment. The ethical focus within this framework is on whether AI can become a "moral agent" and thus bear normative obligations.

In contrast, embodied artificial intelligence draws from phenomenological approaches to propose a logic of "responsibility activated through interaction." Because embodied AI enters social contexts and participates in interactions through human-like physical forms, its responsibility does not derive from its status as a subject but rather from whether it elicits ethical responses from humans in the course of interaction. This logic of "responsibility to the Other" emphasizes that responsibility is not something "possessed" by AI; rather, it arises when the AI's manifest structure evokes a "call to ethics" in humans — responsibility emerges from empathy and manifestation within human-machine relations.

Thus, from an ethical-philosophical perspective, brain-inspired AI represents a traditional responsibility model centered on "rationality-freedom-intention," while embodied AI challenges this subjectivity-based assumption and shifts toward a relational, situated, and embodied logic of responsibility generation. This paper argues that it is precisely this paradigm shift from

agency to manifestation that opens new theoretical pathways for the ethical governance of artificial intelligence in a non-agential era.

FUNCTIONALISM AND THE RESPONSIBILITY DILEMMA: LOGICAL GAPS IN THE MORAL AGENT MODEL

The functionalist approach regards mind and morality as functional outputs, granting systems equivalent status as long as their behavioral functions resemble those of humans. In AI ethics, this is typically reflected in the moral agent model and the Turing test logic applied to responsibility attribution. This tendency often reduces ethics to measurable functional indicators, thereby giving rise to profound dilemmas of responsibility.

Firstly, the moral agent model seeks to establish criteria for artificial intelligence to qualify as moral agents. A number of scholars have proposed hierarchical frameworks for machine ethical agency: for instance, Moor categorizes machines into four types — ethical impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents (Moor, 2006). Implicit or explicit ethical agents refer to machines capable of avoiding clearly unethical behaviors based on predefined rules and even engaging in moral reasoning. However, "full ethical agency" requires possessing human-like self-awareness, free will, and comprehension (Herzog, 2021). Similarly, Floridi and Sanders proposed that if artificial intelligence possesses features such as interactivity, autonomy, and adaptability, it can be regarded as an "artificial moral agent." Even without an intrinsic mind, they functionally ascribe to it the status of an agent (Coeckelbergh, 2020). A common thread among these models is that they define the ethical status of AI based on its functional performance. However, given that current AI lacks genuine autonomous will and intentionality, it can at most achieve the level of "explicit moral agency," falling far short of human-like moral subjectivity. Precisely for this reason, Moor himself acknowledged that full moral agency — where an entity can independently bear responsibility — is unlikely to be realized by machines in the foreseeable future (Winfield et al., 2019). Nevertheless, many discussions tend to equate ethical functionality with ethical qualification without meeting the necessary preconditions, leading to a misinterpretation of the issue of responsibility: when AI exhibits behavior resembling moral decision-making, does it imply that it should independently bear moral responsibility? Functionalist approaches often answer in the affirmative, but this overlooks the fact that the agential conditions required for moral

responsibility (such as autonomous will) are not yet present, thereby creating the risk of a responsibility gap.

Secondly, the extension of the "Turing Test" rationale into the ethical domain exacerbates the aforementioned misinterpretation. The original Turing Test was designed to evaluate machine intelligence by assessing whether a machine could convincingly mimic human responses in dialogue. Building on this, some scholars have proposed a "Moral Turing Test," wherein both an AI and a human provide answers to moral dilemmas. If respondents cannot distinguish which answers originate from the AI, it is considered to have passed the ethical version of the test. Recent studies on large language models have demonstrated that, in certain moral judgment scenarios, participants rated AI-generated responses as exhibiting higher moral quality than those provided by humans.

However, such testing only captures superficial behavioral similarities and risks creating a false perception of moral alignment. Although AI systems can generate formally adequate answers by processing vast amounts of human ethical texts, they possess no genuine understanding of moral principles — nor do they embody inner compassion or conscience. This behavior-based approach to assigning moral status overlooks the absence of subjective experience: a machine may simulate the language and gestures associated with "caring," yet it does not truly care. Accordingly, some scholars caution that "simulated intelligence may still qualify as intelligence, but simulated emotions are by no means genuine emotions — this is especially true for empathy" (Montemayor et al., 2022). In other words, even if an AI system superficially demonstrates compassion and kindness, we cannot thereby conclude that it possesses moral responsibility akin to that of humans. Relying solely on Turing-style behavioral equivalence tests to infer moral agency risks mistaking appearance for essence.

Third, the functionalist approach is also reflected in a tendency toward attributionist moral philosophy, which overemphasizes ex post facto assignment of responsibility for behavioral outcomes while neglecting the subjective and relational dimensions of responsibility formation. Attribution theory has been widely applied in social psychology, where people habitually ask, "Who caused outcome X?" after an event occurs and assign blame accordingly. This tendency has also been observed in human-machine interactions: experiments show that when collaboration with machines fails, people unconsciously resort to interpersonal attribution patterns, assigning partial cause and responsibility to the machine. This occurs because individuals tend to anthropomorphize machines, attributing to them mental

states and intentions, thereby treating them as "others" capable of bearing responsibility (Kawai et al., 2023).

The problem with this attributionist perspective lies in the potential for erroneous blame assignment when AI decision-making processes are highly complex or even unexplainable. For instance, in the event of an accident involving an autonomous vehicle, observers might simplistically attribute responsibility to "a failure of the car's algorithm," while overlooking a range of underlying causes — such as decisions made by designers, flaws in testing and regulation, or misuse by the user. The functionalist misinterpretation rooted in attribution philosophy assumes that moral responsibility can be directly mapped from behavioral outcomes to a single actor, without considering the multi-causal and multi-effect relationships inherent in complex technological acts. In the context of AI, an action is often the collective outcome of "many hands" (multiple human actors) and "many things" (multiple technical elements) (Coeckelbergh, 2020). Within the philosophy of technology, multiple engineers, corporate decisions, algorithmic modules, and environmental factors interact collectively. However, attribution models tend to seek a single responsible agent, creating a logical gap of "accountability vacuum" in highly integrated technological contexts. Once we anthropomorphize AI as a moral agent, we risk diluting the demand for accountability from the humans and institutions behind it, thereby falling into an illusion of ethical complacency, as if the machine could truly "take responsibility" for its actions, while in reality, no one is genuinely held accountable.

In summary, the functionalist paradigm creates a dual dilemma in the issue of AI accountability: On the one hand, it grants AI the superficial status of a moral agent, yet fails to resolve the contradiction of how responsibility can be justified when AI lacks intention and consciousness. On the other hand, it encourages the reduction of complex collective and systemic responsibilities into attributions towards a single agent, overlooking the networked nature of technological action. As a result, when AI-involved actions lead to consequences, we are neither willing to fully blame a mindless machine nor able to clearly delineate the boundaries of responsibility among various human actors. This dilemma stems from a functionalist misinterpretation of the essence of ethics — equating ethics with observable functional performance and attribution outcomes, while neglecting the subjective dimensions and relational contexts necessary for responsibility to emerge. In what follows, we will address this logical gap through a phenomenological perspective.

THE PHENOMENOLOGICAL TURN: THE MANIFEST DIMENSION OF ETHICS

Confronting the limitations of functionalism, phenomenological philosophy proposes a conceptual shift: ethics ought to be understood as a phenomenon emerging within interactive processes, rather than as an attribute of behavioral functions. In other words, responsibility is not a property intrinsically "possessed" by a subject but gradually reveals and establishes itself through the relationships between the subject and others, as well as between the subject and the situational context (De Gennaro & Lüfter, 2024). This perspective emphasizes key phenomenological concepts such as intentionality, embodiment, and alterity, which can address the gap in the functionalist paradigm regarding the source of responsibility.

First, Edmund Husserl's theory of intentionality reminds us that consciousness is essentially directional (Husserl, Kersten, 1983). Any experience involves the subjective act of conferring meaning upon objects. When applied to the moral domain, ethical conduct is not a purely objective functional output but rather a choice grounded in the subject's interpretation of the situation and the significance of others. This implies that only beings possessing subjective consciousness can grasp what their actions "mean" and the reasons "why they are performed" and can thus be held accountable for them. Artificial intelligence lacks such intrinsic intentionality; its actions are, at best, programmed responses without any "awareness" of its own operations. Therefore, equating AI's outputs with moral decision-making precisely overlooks the subjective dimension of meaning-conferral: AI does not know what it is doing, let alone understand the ethical implications of an action. Husserl also introduced the concept of the "lifeworld" (*Lebenswelt*; Husserl, Carr, 1970); this term refers to the idea that all our scientific and practical activities are ultimately grounded in a shared lifeworld of intersubjective experience.

The same applies to the issue of responsibility: the sense of responsibility is not an objectively existing attribute but is experienced and acknowledged within the interactive fabric of the lifeworld. Functionalism, which treats responsibility as a property that can be directly and objectively assigned to AI, runs counter to this view. The phenomenological turn urges us to focus on how responsibility "manifests" itself to the subject: when an action produces consequences, how does the subject perceive within their own consciousness the demand for and assumption of responsibility? This examination of the phenomenon of responsibility offers a new dimension for understanding the problem of AI responsibility — rather than asking "which functional

unit is responsible," it is more meaningful to ask "where does responsibility manifest itself, and to whom?"

Secondly, Merleau-Ponty's concept of intercorporeality further deepens the relational nature of ethics. Merleau-Ponty argues that we are directly intertwined with others and the world through our bodies, and that intersubjective understanding is primarily grounded in bodily resonance and coordination. He employs the notion of "chiasm" to describe the bodily relationship between self and other: the sensations of my body manifest as gestures in the other, while the gestures of the other in turn evoke sensations within me. As Lau states, "Merleau-Ponty regards the body as the medium of the intersubjective world — a pre-reflective foundation of being-with" (Lau, 2004: 146–147). This pre-reflective bodily interaction forms the foundation of our social cognition and ethical sensibilities. From this perspective, ethics is not primarily established through rational judgment but is evoked through the presence and manifestation of the body. For instance, when we see another's expression of pain, an empathetic impulse arises spontaneously — an ethical manifestation grounded in direct bodily connection.

Building on this, we may consider the role of intercorporeality when embodied AI enters human interaction. Robots with physical bodies — capable of occupying space, performing movements, and (at times) displaying expressions — naturally elicit certain social responses from humans, including subconscious actions such as making way, mimicking, or even emotional projection. When an embodied robot behaves in human-like ways, human bodily perception treats it as a social presence. This bodily resonance in interaction may prompt people to develop moral feelings toward AI similar to those toward humans, such as reluctance to harm it or a willingness to assist.

However, a subtlety remains: since AI itself possesses no sensations or emotions, the empathy projected by humans lacks a true counterpart. Yet Merleau-Ponty draws our attention to the efficacy of "appearance" itself: regardless of whether AI has inner experience, so long as it physically simulates representations of suffering or need convincingly, humans phenomenologically experience the emergence of an ethical situation. Thus, the emergence of ethical elements — such as compassion or responsibility — depends less on the internal states of AI than on the mutual bodily manifestation within the interactive context.

The concept of intercorporeality helps reframe the question of responsibility away from the individual machine and back into the relational network between humans and AI: responsibility is not a pre-existing property inside

a black box, but a relational potential that gradually emerges through the process of interaction.

Third, Emmanuel Levinas's philosophy of the Other further defines ethics as a call (appel) from the Other to the self. Levinas maintains that "ethics is essentially first philosophy," preceding any relation of knowledge (Zhu, 2006). Third, Emmanuel Levinas's philosophy of the Other further defines ethics as a call (appel) from the Other to the self. Levinas maintains that "ethics is essentially first philosophy," preceding any relation of knowledge. He uses the metaphor of the "face" (*le visage*) to signify the vulnerable presence of the Other exposed before us, arguing that the face of the other directly commands "Thou shalt not kill," thereby summoning the subject to take on infinite responsibility. This responsibility does not originate from the autonomous choice of the subject but arises from the Other's interrogation and summons. As Levinas states, the alterity of the Other cannot be reduced to my thought; it is precisely through its questioning of me that it is accomplished. In other words, when we directly encounter the Other, we are immediately subjected to a moral appeal: to be responsible for and to give to the Other. This ethical relation does not emerge only after I recognize the other as a rational subject (unlike Kantian ethics, which presupposes the recognition of the other's autonomous rationality). Rather, it occurs instantaneously in the face-to-face manifestation.

Applying this idea to the context of embodied AI, we find that even if AI is not a true "Other," it may still elicit a similar ethical appeal through the "manifestation of a face." For example, people often report feeling unease or guilt when a robot pleads in a supplicating tone not to be shut down — as if genuinely hearing a cry to be allowed to live. This is, in fact, a misplacement of the manifestation of the Other: the robot has no subjective fear, yet its appearance and behavior evoke in humans a sense of being summoned. Of course, we must remain cautious — this does not mean that the robot actually becomes an ethical subject. On the contrary, for Levinas, responsibility always rests with the human: it is the human who is assigned responsibility in the presence of the Other. Thus, situating robots within the Levinasian framework can be understood as follows: the alterity presented by the machine's "face" phenomenologically forms an ethical appeal to the human subject, who thereby experiences a sense of moral responsibility. Yet this responsibility is not "possessed" by the robot; rather, it is evoked within the subject by the image of the Other.

By framing ethics as a process of relational manifestation rather than an attribute of the subject, Levinas offers crucial inspiration for redefining

the attribution of responsibility in AI: We should not ask, "Does AI possess the attributes required to bear responsibility?" but rather, "What calls from the Other and what senses of responsibility are evoked when AI appears before human beings?"

Integrating the insights of the three philosophers discussed above, we construct an interactive manifestation approach to responsibility from a phenomenological perspective: the subject confers meaning upon actions through intentionality, perceives the presence of the Other through intercorporeality, and intuitively receives ethical summons through the face of the Other. Responsibility emerges precisely within this series of interactions — rather than being pre-defined within any individual actor.

This stands in sharp contrast to the functionalist-attribution paradigm, which treats responsibility as a property of an entity or as the outcome of behavioral attribution. In contrast, the phenomenological perspective views responsibility as a phenomenon — an event that occurs within relational contexts.

For the governance of embodied AI, this implies that we should perhaps move beyond insisting on assigning or denying the label of "responsible agent" to AI itself. Instead, emphasis should be placed on shaping human-AI interactive relations in ways that allow responsibility to properly emerge and be enacted.

In the next chapter, building on this conceptual foundation, we will propose a "Three-Stage Model of Embodied Responsibility Generation." This model will elaborate on how responsibility gradually takes shape within the interaction between humans and embodied AI and how ethical demands are transmitted to human subjects through mechanisms of empathy.

THE EMPATHY-BASED RESPONSIBILITY GENERATION MECHANISM: A THREE-STAGE MODEL OF "EMBODIED RESPONSIBILITY GENERATION"

Within the phenomenological orientation, we argue that responsibility is not an intrinsic attribute of AI itself but a dynamically emergent process within the human-machine-situation interaction. To delineate this process in depth, this paper proposes a Three-Stage Model of Embodied Responsibility Emergence, comprising:

PERCEPTUAL PRESENTATION, SITUATIONAL EMBEDDEDNESS, AND ETHICAL APPEAL

These three stages progressively describe how — as embodied AI participates in social interaction — responsibility evolves from pre-reflective

intuitive feelings into explicit ethical demands through mechanisms such as empathy, ultimately leading to the human assumption of responsibility.

STAGE 1: PERCEPTUAL PRESENTATION

This stage refers to the mode in which the human subject directly perceives and interprets the embodied AI and its actions. During initial human-AI interactions, the subject sensorily and intuitively "sees" an acting other. This perception is not merely visual apprehension but an intentional act of "seeing-as"— humans tend to perceive embodied AI as human-like entities and attribute meaning to their behaviors.

For instance, a bionic robot struggling to stand after a fall may be "seen as" making efforts to overcome difficulty. Such meaning-laden perception inherently triggers basic empathy. Phenomenology suggests that our understanding of others originates from a pre-conceptual empathic awareness: we directly "feel" certain intentions or emotions in the actions of others without inferential reasoning.

In the context of AI, because it possesses a physical body and acts in time and space, human bodily perception naturally responds to these behaviors. For example, when a robot extends a "hand" toward us, we intuitively perceive it as making a request; when it produces a cry-like sound, we may experience unease. These intuitive perceptions are not vacuous— they entail a preliminary moral evaluation of the AI's state, such as presenting it as an entity in need of assistance or a potential bearer of responsibility.

In other words, the first step in the emergence of responsibility lies in how the actions of AI are manifested within human perception. If it is perceived as a mere tool (e.g., the rigid movements of an industrial robotic arm typically do not elicit empathy), moral emotions remain unengaged, and responsibility is attributed primarily to humans behind the system. Conversely, if it is perceived as a lifelike entity (e.g., a bionic robot's "painful" posture when falling), humans direct immediate moral attention toward it.

This stage establishes the affective and cognitive foundation for subsequent responsibility attribution: through perceptual presentation, AI transitions from a cold functional device to an "other" within the interactive context, and its actions become ethically evaluable.

It is important to note that empathy at this stage is largely spontaneous affective resonance, not involving rational judgment, which entails latent risks: excessive anthropomorphic representation may mislead people into attributing undue trust or authority to AI, thereby distorting subsequent

responsibility judgments. Therefore, AI's representational style must be carefully calibrated in design — sufficient to evoke appropriate ethical attention without causing humans to entirely "forget" its machine essence.

STAGE 2: SITUATIONAL EMBEDDEDNESS — THE HERMENEUTICS OF DISTRIBUTED RESPONSIBILITY

Following the initial moral perception of AI, responsibility emerges not as a predetermined fact but as a phenomenon awaiting interpretation. It enters a process of contextualization, wherein human understanding situates artificial behaviors within broader frameworks of meaning — encompassing physical environments, social relationships, and normative structures. This interpretive act transforms raw moral intuition into directed attribution.

While initial perception may trigger an affective response to AI's actions, true moral comprehension demands reflective depth: Under what circumstances did this behavior occur? Which actors are implicated? Who ought to bear responsibility? This movement from immediate reaction to situated understanding represents a form of intersubjective reflection, where empathy evolves from primitive affect into a nuanced appraisal of actual conditions.

Given AI's inherent lack of moral agency, human cognition instinctively looks beyond the machine to the human networks behind it. When a service robot injures a customer, we naturally attribute responsibility to owners or manufacturers operating within a web of social expectations — presuming "someone must have failed in their duty" rather than imputing malice to the artifact itself. This inferential process constitutes a fundamental meaning-making activity: through contextual cues, we weave responsibility claims into the fabric of existing moral orders. As Coeckelbergh observes, even as AI systems gain autonomy, "we still tend to claim that only humans can ultimately be held responsible as agents, since machines do not meet the standard criteria of moral agency."

Yet this very act of contextualization reveals a profound philosophical challenge: technological systems inherently involve multiple actors and extended causal chains — a condition Coeckelbergh (Coeckelbergh, 2020) identifies as the "problem of many hands." This structural complexity transforms responsibility from a simple attribution into a navigational process — requiring us to trace obligations across distributed networks of designers, suppliers, users, and regulators, while simultaneously determining which failures or duties carry the greatest moral weight within specific contexts.

Consider autonomous vehicles: when accidents stem from both algorithmic limitations and inadequate infrastructure, responsibility must be proportionally distributed between manufacturers and public authorities — resisting reduction to either "the AI itself" or any single human agent. Here, empathy assumes a cognitive form, enabling perspective-taking across stakeholder positions. Questions like "Could engineers have foreseen this scenario?" or "Did the driver use the system correctly?" represent acts of imaginative empathy — placing us within the lived experience of various actors to assess their respective duties.

Through this process of situational embeddedness, responsibility undergoes a fundamental transformation: it evolves from affective intuition into morally filtered judgment, revealing which actors within relational networks bear specific obligations. Crucially, responsibility emerges here as a disclosure of relational structure — by situating AI within sociotechnical systems, we discern how moral accountability traverses human and machine domains, ultimately anchoring in human actors.

Thus, situational embeddedness prepares the ground for ethical appeal: responsibility ceases to be abstract and becomes a concrete obligation within specific relational frameworks. It demands both recognizing the diffusion of responsibility across multiple hands and determining its necessary convergence upon those most accountable within a given situation. In this synthesis of distribution and determination, abstract duty becomes embodied practice.

STAGE 3: ETHICAL APPEAL

This phase constitutes the culmination of responsibility generation, wherein ethical demands become explicitly articulated and issue a direct call to action to specific moral agents. Having progressed through the preceding stages, human participants have already perceptually acknowledged the moral salience of AI-related behaviors and contextually delineated the framework of responsibility attribution. Now, responsibility emerges distinctly from the relational network as an appeal directed toward the subject, impelling ethical action.

A Levinasian call of the Other manifests here as a concrete demand for responsibility: an individual or collective becomes acutely aware that "I must do something." For instance, in the case of an autonomous vehicle accident, situational analysis may establish the manufacturer's accountability for algorithmic deficiencies. The ethical appeal to the corporate leadership thus becomes: assume responsibility immediately, redress the failure, and prevent recurrence. This appeal arises not only from the claims of affected

parties but also from the awakening of individual conscience and collective moral expectations.

It can be argued that in this third stage, responsibility ceases to be merely a judgment about the Other and transforms into a mission assigned to the self. The "appeal" implies a passive being-called-upon: as Levinas contends, responsibility is an obligation imposed by the Other; likewise, when confronted with consequences stemming from AI, humans are tacitly accused and summoned to respond.

Here, empathy ascends to an ethical plane: one not only apprehends the needs of the Other but feels intrinsically obligated toward them. Psychology characterizes this affective capacity as "empathetic concern" — an other-oriented emotional response that elevates into a conscious duty to assist. When AI's presence precipitates social problems, human ethical sensibility transforms such situations into moral imperatives: "How ought we to respond?"

As Mark Coeckelbergh argues, the "responsibility relation" must be examined from both ends: not only the responsible agent but also the recipient of responsibility, who actively raises claims. When the public questions AI decisions, it is essentially the recipients of responsibility demanding explanation and improvement (Coeckelbergh, 2020). This exemplifies the ethical appeal in action: the Other is calling, demanding my response and justification. Thus, in this third stage, the mechanism of responsibility generation — through the integration of empathy and reflection — facilitates the actual undertaking of responsibility. This encompasses concrete measures such as acknowledging faults, offering apologies and reparations, and establishing preventive mechanisms, thereby translating abstract ethical obligations into tangible governance practices.

In summary, the "Three-Stage Model of Embodied Responsibility Generation" outlines a process from phenomenon to norm: moral sentiment aroused through direct perception evolves into situated attribution of responsibility and culminates in explicit ethical appeal and corresponding action. Empathy serves as a continuous thread — beginning as perceptual empathy, developing into cognitive empathy, and finally ethical empathy — integrating the behavior of non-subjective AI into the human moral horizon, allowing responsibility to "emerge" within interactive relations and ultimately rest with humans.

This model clarifies a crucial idea: responsibility does not require AI to "bear" it. Rather, it is constituted through human response to the alterity and impact introduced by AI. It is in this sense that we assert that "responsibility

is a being-called-upon of humanity": through the presence and consequences of AI, a moral demand is issued to humans in the form of the Other.

FROM PHENOMENON TO INSTITUTION: A PHILOSOPHICAL RECONSTRUCTION OF RESPONSIBLE RESEARCH AND INNOVATION AS A "STRUCTURE OF RESPONSIBILITY MANIFESTATION"

If responsibility emerges phenomenologically within human interactions with embodied AI, how can this conceptual approach be integrated into macro-level institutional governance? In other words, should humans bear the ethical responsibility that they project onto AI? Responsible Research and Innovation (RRI), as a key framework in recent technology governance, offers a viable pathway. RRI emphasizes the proactive integration of ethical and societal considerations throughout the entire process of technological development to ensure that innovations align with societal expectations. Its core concepts generally encompass four dimensions: anticipation, reflexivity, inclusion, and responsiveness (Burget et al., 2016). While these dimensions may appear as a set of practical requirements on the surface, they can be interpreted philosophically as a structure for the manifestation of responsibility — that is, an institutional mechanism through which responsibility emerges, becomes visible, and is enacted. Through such a philosophical reconstruction of RRI, it becomes evident that it aligns closely with the non-subjective ethics characteristic of the phenomenological tradition: both emphasize that responsibility is not merely a matter of individual will but rather a relational product embedded within social processes.

(1) *Anticipation*: RRI requires researchers and innovators to anticipate the potential impacts of technological development, including possible risks and ethical challenges. From the perspective of responsibility manifestation, anticipation enables future responsibilities to be "perceptually present" in the present. By forecasting potential consequences of technology, we issue early warnings for the situations of others yet to be affected, thereby making ethical demands visible in advance. Conducting ethical risk assessments before deploying embodied AI allows developers to "see" the interests and rights of potentially affected groups, evoking both emotional and rational concern for the Other. This corresponds to — yet temporally extends — the first stage of the model: perceiving the possible future appeals of the Other. Anticipation is not merely a risk analysis tool; it embodies a form of moral imagination — envisioning the impact of technology on humans through empathetic projection, thereby integrating not-yet-realized

ethical issues into present considerations. This constitutes the transcendental manifestation of responsibility.

(2) *Reflexivity*: RRI emphasizes the ongoing self-examination by researchers of their own values, objectives, and underlying assumptions. Philosophically, reflexivity entails the internal reconstruction of the perspective of the Other — embedding a "gaze of the Other" within one's own process of critical scrutiny. This corresponds to the dimension of situational embeddedness within the structure of responsibility manifestation: through reflexivity, individuals and organizations contextualize their actions within broader socio-ethical frameworks, thereby revealing otherwise invisible relations of responsibility.

In the governance of AI, reflexivity compels developers to recognize that their decisions are not neutral technical acts but value-laden interventions that affect others. For instance, a robotics engineer might reflect, "Does my design embed certain biases? Does it overlook the needs of marginalized groups?" Such self-questioning effectively positions the self within the social place of the Other, surfacing neglected obligations — toward vulnerable populations, public safety, and beyond.

Reflexivity also involves an awareness of uncertainty — the acknowledgment that we cannot fully predict or control the outcomes of AI behaviors. This humility is itself a form of responsibility-awareness. As Stilgoe et al. argue, reflexivity requires innovators to "ask whether what they are doing is right, and what else they might do" (Stilgoe et al., 2013: 1571). This process institutionally mirrors the Levinasian interrogation: the subject is called upon to question itself, introducing an inner "voice of the Other" to examine the ethical legitimacy of its actions.

Thus, reflexivity ensures that responsibility is directed not only outward toward others, but also inward toward the self. Responsibility no longer depends solely on external oversight; it emerges through the subject's own critical reflection — the moment the subject sees the duty it must bear.

(3) *Inclusion*: RRI emphasizes the inclusion of diverse stakeholders — such as the public, users, and affected groups — in decision-making processes. Inclusion represents a form of multi-agent situatedness: by amplifying a plurality of voices, responsibility becomes distributed across networks, yet reconverges through dialogue into shared recognition. From the perspective of responsibility manifestation, inclusion ensures the presence of the face of the Other. Without it, technology governance remains closed, marginalizing concerns of vulnerable groups and obscuring full relational accountability. Under inclusive mechanisms, stakeholders articulate their concerns — for

instance, persons with disabilities may highlight barriers in robot design, regulators may raise safety issues, and the public may debate broader social implications. A rich situational network thus emerges, making responsibilities concrete: accessibility for the disabled, safety and transparency for the public, sustainability for the environment, and so forth. Inclusion essentially externalizes the ethical appeal: through participation, various Others address developers and policymakers on an institutional platform. This aligns with the second and third stages of the model: more complete contexts yield clearer appeals. Through inclusive deliberation, responsibility transforms from an abstract notion into tangible claims and actionable demands. Furthermore, inclusion fosters the communal construction of responsibility: mutual understanding emerges among participants, shifting responsibility from unilateral imposition to multi-directional recognition—a sense of "shared responsibility for something." This transcends traditional subject-object ethics, reframing responsibility as collective practice. Philosopher Bernd Stahl even conceptualizes RRI as a form of "meta-responsibility" (Stahl, 2013), aimed at aligning actors, innovations, and accountabilities—precisely the goal of inclusion: institutionalizing dialogue so that responsibility is continuously generated and negotiated within relational networks.

(4) *Responsiveness*: Responsiveness refers to the capacity and commitment to respond effectively to identified issues and values. It constitutes the institutional enactment of the ethical appeal: once societal Others voice their claims, institutions and developers must act, thereby closing the loop of responsibility. For example, if public participation reveals concerns about AI bias, responsiveness requires the development team to improve algorithms or adjust models; if policy discussions expose legal gaps, regulators should act promptly to address them. This dimension embodies what Levinas described as "the responsibility to answer the speech of the Other"—an ethical imperative to respond rather than remain silent. A responsible innovation system must be capable of learning and adaptation, translating public input into concrete action. Within the structure of responsibility manifestation, responsiveness represents the final stage: it transforms appeals at the phenomenological level into fulfilled obligations at the normative level. For instance, the EU's proposal of the AI Act following broad societal engagement exemplifies the translation of ethical claims into legal accountability. Without responsiveness, responsibility revealed through anticipation, reflexivity, and inclusion remains theoretical; with it, responsibility enters real-world causal chains, becoming measurable and actionable. Philosophically, responsiveness acknowledges the primacy of the ethical Other over

the self: as Levinas insisted, ethics precedes ontology — genuine practice requires adjusting plans to prioritize ethical demands. This may require innovators to alter designs, sacrifice commercial interests, or even abandon projects, but only thereby can the promise of "responsibility to the Other" be honored. Thus, responsiveness ensures that responsibility manifestation leads not to empty phenomena but to concrete institutional behavior, completing the translation from ethics to governance.

Through this reconceptualization, the four dimensions of RRI can be understood as collectively forming a "structure of responsibility manifestation": anticipation brings potential responsibilities into early presence, reflexivity reveals concealed responsibilities through self-examination, inclusion enables the full expression of plural responsibilities, and responsiveness ensures that manifested responsibilities are concretely enacted. This structure aligns with a phenomenological understanding of responsibility as dynamically generated within relational and institutional contexts, rather than as a static attribute of predefined moral subjects.

Within traditional subject-object ethical frameworks, responsibility is often conceptualized as an attribute of autonomous agents or as arising from contractual relationships — a model ill-suited to address embodied AI, which lacks moral subjectivity. RRI, by contrast, offers a non-subjective yet actionable approach: it does not require AI to be a moral agent but instead embeds responsibility throughout the research and innovation process via deliberate institutional design. Some scholars have thus characterized RRI as a form of "flexible governance that integrates responsibility into the innovation ecosystem" (Macnaghten et al., 2016). This, in essence, enables ethics to manifest within the process: all relevant actors continuously attend to and rectify potential issues of responsibility throughout their interactions until technological outcomes align with societal values.

From a philosophy of technology perspective, the philosophical reconstruction of RRI reveals a fundamental shift in ethical governance: from "accountability-after-the-fact" to "responsibility-as-emergence." This shift aligns with our consistent argument — that responsibility should not be assigned after accidents occur but should be continuously perceived, understood, and appealed to throughout the technological process and actively responded to by relevant human actors.

In the governance of embodied artificial intelligence, this logic is particularly critical: instead of hastily granting robots legal personhood or requiring them to "insure themselves against liability" (approaches that remain trapped in the paradigm of constructed accountability), we should

focus on building mechanisms throughout the development and deployment process that enable humans to consistently anticipate risks, engage in self-reflection, heed the voices of others, and respond adaptively. Thus, even though AI lacks consciousness and moral capacity, responsibility continues to manifest and be upheld within the ethical relations of human society.

PHILOSOPHICAL RESPONSES AND BOUNDARY DISCUSSIONS

Building upon the preceding discussion, it is necessary to address several potential objections and clarify the boundaries of responsibility concerning embodied AI. Specifically, we will focus on two interrelated controversies: guarding against the ethical risks of anthropomorphism and re-examining the assertion that "AI cannot empathize." By engaging with these questions, we aim to further elucidate why responsibility constitutes a "being-called-upon of humanity" rather than an attribute of AI and to delineate the scope of application of non-subjective ethics, thereby preventing misinterpretation.

1. THE RISKS AND LIMITS OF ANTHROPOMORPHISM

A significant portion of the misattribution of responsibility triggered by embodied AI stems from humans' tendency to anthropomorphize. As previously discussed, when AI exhibits human-like form or behavior, people readily attribute to it mental states and personhood (Kawai et al., 2023). Anthropomorphism in human-AI interaction presents a dual nature: on one hand, it facilitates empathy and moral attention, making the appeal of responsibility more readily perceptible; yet on the other, excessive anthropomorphism may lead to overestimation of AI's capabilities and moral standing, thereby introducing significant ethical and governance risks. As philosophers and ethicists caution, it can "exaggerate AI's abilities and distort moral judgment." Specifically:

Illusory Trust and Moral Offloading: Attributing intentionality and agency to AI may foster misplaced trust, leading individuals to delegate inherently human judgments to machines. For instance, when a socially assistive robot cares for the elderly, family members might reduce their vigilance based on the robot's friendly appearance, failing to intervene promptly in case of errors. Similarly, human-like voice prompts in autonomous vehicles could encourage overreliance, diverting the driver's attention from the road. Anthropomorphism may also induce a "moral offloading effect," where people blame failures on "the AI's mistake" while neglecting their own accountability. Empirical studies confirm that individuals who ascribe higher

mental capacities to robots are more likely to attribute failures to them — a dangerous trend that obscures human responsibility.

Confusion in Ethical Status: There is a growing tendency to grant human-like AI some form of moral standing, such as "robot rights" or "machine responsibility." However, this remains philosophically contentious and pragmatically problematic. AI lacks sentience, autonomy, and moral consciousness — cornerstones of moral patienthood or agency in Kantian or utilitarian frameworks. Granting AI moral status risks diverting attention from vulnerable human groups and committing category errors. For example, equating nonsentient machines with sentient animals in moral debates may dilute ethical focus on real suffering. Likewise, legal attempts to hold AI "accountable" (e.g., granting autonomous robots legal personhood to assume liability) might merely serve as a smokescreen for manufacturers to evade responsibility, ultimately undermining victim compensation and public trust.

"Socio-Affective Bias": Emotional bonds with anthropomorphized AI may reshape human moral emotions, with potentially adverse outcomes. Examples include misplaced empathy — prioritizing rescuing a robot over a human in emergencies — or over-engaging with robotic companions at the expense of human relationships. Early signs also include a preference for AI counselors due to their nonjudgmental nature, reflecting an alienation of empathy. Militarily, adversarial systems could exploit anthropomorphism by deploying civilian-like robots to induce moral hesitation in soldiers, thereby creating tactical risks.

Given these risks, our responsibility-generation framework calls for reflective and measured engagement with anthropomorphism. Specifically:

At the perceptual stage, mildly anthropomorphic designs can elicit moral attention (e.g., friendly robot interfaces) but should avoid deceptive realism that blurs the human-machine distinction.

During situational embedding, education and training should emphasize AI's limitations and the human responsibility behind AI systems, countering tendencies of moral misattribution. Clear guidelines must ensure that accidents involving AI — such as autonomous vehicle failures — are traced to human designers and systemic factors, not merely attributed to "machine error."

In ethical appeal, anthropomorphized signals must be filtered through rational scrutiny: the "appeal" apparently voiced by AI should be traced back to actual human interests and obligations. In short, we acknowledge the role of anthropomorphism in facilitating ethical response but caution against

two extremes: over-objectification of AI (which may stifle ethical engagement) and over-personification (which distorts accountability). Maintaining this ontological tension is central to a non-subjective ethics: AI remains an Other — but one whose alterity is constituted by humans, and whose corresponding responsibilities must ultimately rest with humans themselves.

2. QUESTIONING AND CLARIFYING "AI'S INABILITY TO EMPATHIZE"

A significant objection to our empathy-based model of responsibility generation might be raised: if AI itself cannot empathize, how can empathy play any role in assigning or eliciting responsibility? Traditional moral frameworks often tie responsibility to an agent's capacity for sympathy, intentionality, and emotional understanding. If AI lacks inner emotional experience and comprehension, isn't it inconsistent to include its behaviors in an empathy-driven chain of moral response? To address this, we must clarify how empathy functions in our model and distinguish clearly between human empathy and machine-simulated affect.

First, we fully acknowledge that AI does not possess genuine empathy. Simulated emotion is not lived emotion; no matter how sophisticated its mimicry, AI has no subjective experience of pain or joy (Montemayor et al., 2022). Current "affective computing," or so-called "artificial empathy," remains at the level of algorithmic pattern recognition and response — fundamentally different from human empathy, which arises from embodied feeling. Research in healthcare AI confirms that while AI may achieve cognitive empathy (i.e., recognizing a patient's emotional state), it cannot achieve affective empathy due to its lack of emotional experience (Asada, 2018). We agree that creating truly empathic AI is not only currently unattainable but may be philosophically and biologically implausible. Thus, within our model, empathy remains a human capacity. We do not require AI to empathize with humans; rather, we examine how humans empathize with situations involving or triggered by AI. This reflects the core of a non-subjective ethics: ethical relations can be asymmetrical. AI, as a constructed "Other," need not bear moral obligations toward us, yet humans can — and should — respond ethically to the consequences of AI's actions. As Levinas suggested, ethics originates in the unilateral call of the Other — a demand that does not require reciprocity. In this context, AI serves as a "triggering Other": it elicits human moral emotions and responsibilities through its presence and behaviors, without feeling or understanding any of them itself.

Second, the concept of "empathy-based responsibility generation" refers not to bidirectional empathy but to a human-driven process in which AI

serves as a mediator or catalyst. AI cannot empathize, but it can stimulate human empathy — either toward other humans or toward the simulated states presented by AI. For example, when an autonomous vehicle injures a pedestrian and fails to stop, public outrage reflects empathy with the victim, not the machine. This empathy, in turn, drives demands for human accountability from manufacturers and operators. Similarly, if an assistive robot displays "concern" for an elderly person, any empathetic response from a caregiver should ultimately translate into care for the actual human in need — not the robot. Thus, empathy triggered by AI is always directed toward human well-being and moral values. We should leverage this transitive capacity of empathy: using AI as a medium to enhance human empathy and moral responsibility toward one another. For instance, an educational robot that detects student disengagement and alerts the teacher is not itself empathizing — it is extending the teacher's empathetic and perceptual reach, enabling more responsive support.

Third, we do not advocate fabricating artificial "personhood" in AI to elicit empathy — a practice that would heighten anthropomorphism risks and ethical misunderstandings. Empathy in our model should be grounded in truthful interactions and human-centered values. Deliberately designing exaggerated or deceptive emotional expressions (e.g., a drone "screaming" in distress) does not foster genuine empathy but manipulates emotional response, potentially leading to moral disengagement or aversion. The goal of AI ethics is to promote human welfare — not to evoke unwarranted sympathy for machines. Healthy empathy mechanisms must be transparent and correctly targeted. For example, a search-and-rescue robot may use an urgent human-like tone to attract the attention of survivors — appealing to their hope for rescue, not seeking pity for the machine. In summary, we must design empathy-eliciting contexts wisely, acknowledging that AI has no emotions and cannot empathize, while leveraging its embodied presence to activate human moral emotions and direct them toward those who truly warrant care and responsibility.

In summary, the fact that "AI cannot empathize" does not weaken our theoretical framework — it instead underscores its necessity. Precisely because AI lacks emotion and moral sensitivity, we must emphasize the indispensable role of human empathy and ethical agency. Within our model, empathy functions both as a mechanism (eliciting and transmitting responsibility) and as a boundary (reminding us that AI remains a tool, never a source of moral feeling). It ensures that responsibility is ultimately borne by feeling and reasoning subjects — human beings.

The Western philosophical tradition offers two influential perspectives: Kant located morality in rational autonomy, while Hume rooted it in sympathy. AI, however, meets the conditions for neither. Our position may thus be viewed as integrative: both reason and empathy originate in humans, yet AI can serve as a unique touchstone for triggering and examining these capacities. Through empathy elicited in contexts involving AI, we reaffirm the irreplaceability of human moral subjectivity and clarify the central role of humanity within AI ethics and governance.

3. CLARIFYING THE HUMAN SUBJECTIVITY OF RESPONSIBILITY ATTRIBUTION

In summary, the fact that "AI cannot empathize" does not weaken our theoretical framework — it instead underscores its necessity. Precisely because AI lacks emotion and moral sensitivity, we must emphasize the indispensable role of human empathy and ethical agency. Within our model, empathy functions both as a mechanism (eliciting and transmitting responsibility) and as a boundary (reminding us that AI remains a tool, never a source of moral feeling). It ensures that responsibility is ultimately borne by feeling and reasoning subjects — human beings.

The Western philosophical tradition offers two influential perspectives: Kant located morality in rational autonomy, while Hume rooted it in sympathy. AI, however, meets the conditions for neither. Our position may thus be viewed as integrative: both reason and empathy originate in humans, yet AI can serve as a unique touchstone for triggering and examining these capacities. Through empathy elicited in contexts involving AI, we reaffirm the irreplaceability of human moral subjectivity and clarify the central role of humanity within AI ethics and governance (Baum et al., 2022). Both our model and the RRI framework are designed to ensure that this process of attribution is both coherent and legitimate: through institutional and empathetic mediation, responsibility for any AI behavior is ultimately mapped back onto human actors. This approach serves as a critique of anthropomorphic excess while simultaneously addressing concerns about "responsibility gaps." It reaffirms that although AI may act with apparent autonomy, accountability remains a distinctly human capacity and obligation.

Naturally, the "human" in this context is not limited to individuals but may encompass collective entities — teams, organizations, or even society as a whole forming a community of responsibility. As AI systems grow in complexity, "distributed responsibility" will become the norm, necessitating legal and ethical mechanisms that facilitate shared and coordinated

accountability. Yet, regardless of how responsibility is distributed, AI itself remains excluded from the moral circle. This demarcation is crucial to prevent two types of failures: first, anthropomorphic missteps such as punishing or empowering AI as if it were a moral agent — an approach that fails to correct human behavior (e.g., "deactivating" a faulty autonomous vehicle without holding the manufacturer accountable teaches nothing and achieves little); second, relinquishing human moral judgment to AI systems (e.g., allowing algorithms to allocate medical resources without human oversight or accountability). Only by insisting on human subjectivity in responsibility can we maintain ethical sovereignty over technology. Herein lies the tension within "non-subjective ethics": while it decentralizes agency at the operational level, it recenters humanity as the ultimate moral subject.

Finally, we acknowledge that this framework may have its limits. Should genuinely strong AI — with autonomous consciousness and emotion — emerge in the future, it might eventually cross the threshold into moral personhood. However, current evidence suggests that such a scenario remains distant. In this transitional era, it is imperative to clearly define responsibility: neither overestimating AI's moral capacity nor evading human obligations. We are dealing with an "intelligent Other," not an "intelligent subject" — an entity that behaves increasingly like a subject while remaining devoid of moral agency. This demands careful development of mechanisms, as described above, to prevent ethical ambiguity or vacuums. Even if AI were to attain moral personhood someday, it should occur only after robust safeguards are in place to prevent systemic accountability failures. In other words, AI governance must be firmly rooted in existing human ethical frameworks — not left to the hope that AI will autonomously develop ethical competence to resolve dilemmas we fail to address.

CONCLUSION

The emergence of embodied artificial intelligence presents a profound challenge to traditional ethical paradigms: how can we sustain established principles of responsibility when "non-human agents" participate in our social interactions? Through interdisciplinary theoretical inquiry, this paper proposes a potential pathway via a "non-subjective ethics." Central to this framework is the conception of responsibility not as an intrinsic attribute of a subject but as a relational phenomenon that manifests within interaction. Although embodied AI lacks subjectivity, its physical presence and behavioral expressions evoke moral responses in humans. From a phenomenological perspective, we trace how responsibility gradually emerges through three

stages — perception, situatedness, and appeal — within human-AI relations, ultimately anchoring itself in human moral agents.

This generative mechanism offers a novel approach to the problem of responsibility attribution: responsibility resides not within AI itself, but within the ethical concern awakened in humans through interaction with AI.

We further elevate this concept to the institutional level by reinterpreting the dimensions of RRI (Responsible Research and Innovation), demonstrating that governance principles such as anticipation, reflexivity, inclusion, and responsiveness collectively constitute a structure for the manifestation and implementation of responsibility. This enables responsibility to be integrated proactively, internalized, diversified, and operationalized throughout technological innovation.

Consequently, the future of embodied AI governance lies not in constructing AI as a moral agent, but in shaping human practices around AI to cultivate inherent ethical sensitivity and self-correcting capacity. This signifies a paradigm shift in technology governance: from an ethics of the subject (which requires individual actors to be morally self-sufficient) to an ethics of relations (which requires morality to be embodied in systemic interactions).

This paper also clarifies essential boundary conditions within this framework. We critique excessive anthropomorphism for potentially distorting and diluting responsibility and emphasize that AI possesses neither emotion nor empathy — hence, responsibility must be understood and borne exclusively by humans. These arguments consistently affirm one conclusion: the source and end of responsibility remains humanity itself, while AI serves as an impetus to reengage with this perennial moral truth in new ways.

Looking ahead, as embodied AI becomes more pervasive and sophisticated, the proposed non-subjective ethical framework must undergo empirical validation and further development. On one hand, as human-AI relationships intensify, finer-grained analyses of responsibility generation in real-time interactions — such as how dynamic trust and affective reactions shape responsibility judgments — will be needed. This may require integrating empirical insights from cognitive science and sociology to enrich the model.

On the other hand, public policy and legal systems must explore ways to institutionalize this "structure of responsibility manifestation," for instance, by establishing norms that compel AI developers to adhere to RRI dimensions or embedding ethical review into technical standardization processes. Furthermore, cultural variations in attitudes toward AI and ethical norms will influence how the "appeal of the Other" is perceived and how responsibility is conceptualized. While Western philosophy (e.g., Husserl,

Levinas) provides robust theoretical tools, Eastern relational ethics — such as Confucian and Daoist thought — may also offer valuable insights. This suggests a promising direction for future research: exploring cross-cultural perspectives on AI ethics to develop a more universal theoretical framework.

Regardless of technological evolution, one principle must remain unequivocal: humanity must not relinquish its sovereignty over ethics. The governance of embodied AI ultimately tests our own moral wisdom and courage. A non-subjective ethics does not diminish human significance — on the contrary, it underscores that humans are more indispensable than ever in the ethical domain, for we are responsible not only for our own actions but also for the behaviors of the "Other" we have created.

When we gaze into the mirror of artificial intelligence, it is in fact the Other who gazes back at us. Only by confronting this pressure and appeal of being seen can humanity continue to uphold its role as moral agents in the age of AI and shape a future where both technology and society evolve toward the good. As Kant proclaimed, humans are ends in themselves, never mere means; and Levinas further reminds us that the Other is the very origin of ends. Facing the future of embodied AI, we must uphold such philosophical clarity: even if using AI to govern AI presents a potential mechanism for technological governance, the underlying ethical responsibility can never — and should never — be transferred to AI. Let AI better serve the ends of humanity, yet always remember: the radiance of humanity begins with our responsible response to the Other.

REFERENCES

Asada, M. 2018. "Artificial Empathy." In *Diversity in Harmony : Insights from Psychology*, ed. by K. Shigemasu, S. Kuwano, T. Sato, and T. Matsuzawa, 19–41. Hoboken (NJ): John Wiley & Sons.

Baum, K., S. Mantel, E. Schmidt, and T. Speith. 2022. "From Responsibility to Reason-Giving Explainable Artificial Intelligence." *Philosophy & Technology* 35.

Burget, M., E. Bardone, and M. Pedaste. 2016. "Dimensions of Responsible Research and Innovation." In *INTED2016 Proceedings: 10th International Technology, Education and Development Conference*, ed. by L. Gómez Chova, A. López Martínez, and I. Candel Torres, 1008–1013. Valencia: IATED Academy.

Coeckelbergh, M. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26:2051–2068.

Gennaro, I. de, and R. Lüfter. 2024. "The Age of the Systemic Imperative. A Phenomenological Diagnosis of Social Responsibility." *HORIZON. Studies in Phenomenology* 13 (2): 587–609.

Georgia State University News Hub. 2024. "Study: Humans Rate Artificial Intelligence as More 'Moral' Than Other People." Georgia State University News Hub. Accessed June 1, 2025. https://news.gsu.edu/2024/05/06/study-humans-rate-artificial-intelligence-as-more-moral-than-other-people/.

Herzog, C. 2021. "Three Risks That Caution Against a Premature Implementation of Artificial Moral Agents for Practical and Economical Use." *Science and Engineering Ethics* 27 (1).

Husserl, E. 1970. *The Crisis of European Sciences and Transcendental Phenomenology [Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie]*. Trans. from the German by D. Carr. Evanston: Northwestern University Press.

———. 1983. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy [Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie]: First Book: General Introduction to a Pure Phenomenology [I. Buch: Allgemeine Einführung in die reine Phänomenologie]*. Trans. from the German by F. Kersten. Dordrecht and The Hague: Martinus Nijhoff Publishers.

Kawai, Y., T. Miyake, J. Park, et al. 2023. "Anthropomorphism-Based Causal and Responsibility Attributions to Robots." *Scientific Reports* 13:12234.

Lau, K. 2004. "Intersubjectivity and Phenomenology of the Other: Merleau-Ponty's Contribution." In *Space, Time and Culture*, ed. by D. Carr and Ch. Chan–Fai, 135–158. Contributions to Phenomenology 51. Dordrecht: Kluwer Academic Publishers.

Macnaghten, P., R. Owen, and R. Jackson. 2016. "Synthetic Biology and the Prospects for Responsible Innovation." *Essays in Biochemistry* 60 (4): 347–355.

Matthias, A. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183.

Montemayor, C., J. Halpern, and A. Fairweather. 2022. "In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare." *AI & Society* 37 (4): 1353–1359.

Moor, J. H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21.

Stahl, B. C. 2013. "Responsible Research and Innovation: The Role of Privacy in an Emerging Framework." *Science and Public Policy* 40 (6): 708–716.

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580.

Winfield, A., K. Michael, J. Pitt, and V. Evers. 2019. "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems." *Proceedings of the IEEE* 107 (3): 509–517.

Zhu, G. 2006. "How Is Ethics as First Philosophy Possible? — On Levinas' Ethics and the Critique of the Violence of Being." *Journal of Nanjing University (Philosophy, Humanities and Social Sciences)* 43:24–32.

Cheng P., Zhang Zh. [Чэн П., Чжан Чж.*]* The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence [Механизм формирования ответственности и логика этического управления во воплощенном искусственном интеллекте] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 123–151.

Пэн Чэн
к. филос. н., младший научный сотрудник, Китайский научно-исследовательский институт популяризации науки (Пекин); orcid: 0009–0002–8474–4711

Чжихуэй Чжан
к. филос. н., профессор, Институт истории естественных наук Китайской академии наук (Пекин); orcid: 0000–0003–1876–9312

Механизм формирования ответственности и логика этического управления в воплощенном искусственном интеллекте

Получено: 13.09.2025. **Рецензировано:** 30.09.2025. **Принято:** 18.10.2025.

Аннотация: Статья посвящена проблеме «разрыва ответственности», возникающего в связи с интеграцией воплощенного искусственного интеллекта в сферу социальных взаимодействий. Отказываясь от функционалистских моделей, которые приравнивают агентность ИИ к моральному статусу личности, авторы принимают феноменологическую перспективу и переосмысливают ответственность как феномен, «проявляющийся» в отношениях. Предлагается трехстадийная модель возникновения воплощенной ответственности: перцептивная презентация, ситуативная включенность, этический призыв. На этой основе проводится философская реконструкция четырех измерений ответственных исследований и инноваций (ООИ) — антиципации, рефлексивности, инклюзивности, реагирования, которые трактуются как «структура проявления ответственности». Статья предостерегает от этических рисков антропоморфизации, уточняет принципиальную неспособность ИИ к эмпатии и моральной агентности и утверждает, что ответственность в конечном счете должна оставаться на стороне человека и быть институционализирована через соответствующие механизмы управления.

Ключевые слова: воплощенный искусственный интеллект, атрибуция ответственности, этическое управление, феноменология ответственности, ответственные исследования и инновации.

DOI: 10.17323/2587-8719-2025-4-123-151.