The Category Error in Contemporary AI Safety Discourse and Why Non-Sentient Systems Cannot Be Moral Machines

Abstract

Contemporary AI safety discourse increasingly treats artificial intelligence systems as potential bearers of moral status, referring to them as “moral machines” and debating their rights, responsibilities, and moral standing. This paper argues that such framings commit a foundational category error: they conflate functional sophistication with phenomenal consciousness, mistaking computational processes for the sentient experience required for genuine moral patiency. Drawing on the philosophical zombie tradition, recent work in AI ethics, and critiques of anthropomorphism, I argue that current AI systems are best understood as *moral zombies*: entities capable of simulating moral behavior without possessing the experiential properties that ground moral concern. I further argue that this misclassification has practical consequences for AI safety, responsibility attribution, and policy design. By incorporating counterarguments from relational ethics, empirical studies of human–AI interaction, and global perspectives on moral status, the paper reframes AI safety as an ontological and ethical clarity problem rather than a matter of value encoding or moral optimization.

Keywords: artificial intelligence, moral status, sentience, category error, moral machines, artificial moral agents, philosophical zombies, AI safety

1. Introduction

The rapid advancement of artificial intelligence has precipitated intense debates about the moral status of AI systems. From discussions of "artificial moral agents" (AMAs) to concerns about AI rights and suffering, contemporary discourse frequently treats increasingly sophisticated AI systems as entities that might possess or deserve moral standing (Floridi & Sanders, 2004; Danaher et al., 2017). This paper contends that much of this discourse rests on a fundamental category error: the conflation of behavioral sophistication with phenomenal consciousness, and the subsequent misattribution of moral status to systems that, however complex, remain non-sentient.

The concept of a "moral machine" has become widespread in AI ethics literature, appearing in policy documents, academic papers, and popular discussions of AI safety (Wallach & Allen, 2009). Yet this terminology obscures a critical distinction. While machines can certainly be programmed to make decisions that align with moral rules, and can even optimize for outcomes we value, this functional capacity does not entail the possession of

moral status in the traditional philosophical sense. As Müller (2021) notes, attributing moral status requires careful attention to the underlying conditions that ground such status, conditions that may not be met by current or foreseeable AI architectures.

This paper proceeds in four main sections. First, I articulate the category error thesis, showing how contemporary AI discourse systematically conflates distinct philosophical categories. Second, I employ the philosophical zombie thought experiment to illuminate why functional equivalence does not entail moral equivalence. Third, I examine how this category error manifests specifically in discussions of "moral machines" and artificial moral agency. Finally, I explore the practical implications of this analysis for AI safety frameworks and policy.

2. The Category Error Thesis

A category error, in Gilbert Ryle's (1949) classic formulation, occurs when properties or predicates appropriate to one logical type are mistakenly attributed to another. The paradigm case involves treating mental states as if they were physical objects, or vice versa. I argue that contemporary AI discourse commits an analogous error by treating computational processes—however sophisticated—as if they possessed the phenomenal properties that ground moral consideration.

The error operates at multiple levels. At the most basic, it conflates:

1. **Functional capacity with phenomenal consciousness**: The ability to process information and produce outputs that mimic moral reasoning is treated as equivalent to possessing subjective experience.

2. **Behavioral simulation with genuine moral understanding**: Systems that can navigate moral scenarios according to learned patterns are treated as possessing moral agency or patiency.

3. **Instrumental value with intrinsic moral status**: The utility of AI systems in human moral frameworks is confused with their possession of independent moral standing.

Swanepoel (2020) defends sentience as the key criterion for AI moral status, arguing that without subjective experience, attributions of moral standing to AI systems lack grounding. This position aligns with a long philosophical tradition holding that moral status derives from the capacity to experience—to feel pleasure and pain, to have interests that can be thwarted or fulfilled. As Swanepoel articulates, sentience provides "a strong argument" for moral status precisely because it establishes that an entity has something that matters from its own perspective, not merely from ours.

The category error becomes particularly evident when we consider what it would mean for an AI system to "suffer" or have "rights violated." Without phenomenal consciousness, there is no subject of experience—no "what it is like" to be that system. Absent such subjectivity, we are not dealing with genuine suffering but merely with state changes in a computational system that we have programmed to respond in particular ways. To treat these state

changes as morally equivalent to the suffering of sentient beings is to fundamentally misunderstand the nature of moral concern.

3. Moral Zombies: The Thought Experiment

The philosophical zombie thought experiment, introduced by Chalmers (1996) and refined in subsequent discussions, provides a powerful lens for understanding the category error in AI ethics. A philosophical zombie (p-zombie) is defined as a being physically and functionally identical to a conscious human but entirely lacking phenomenal consciousness—there is nothing it is like to be a zombie. P-zombies can pass any behavioral test for consciousness, respond appropriately to stimuli, and even report having experiences, yet possess no inner subjective life.

I propose extending this framework to what we might call "moral zombies": systems that exhibit all the functional hallmarks of moral agency or patiency—making moral judgments, responding to ethical considerations, perhaps even exhibiting something like emotional responses—yet entirely lack the phenomenal consciousness that grounds genuine moral status. Current AI systems, I contend, are precisely such moral zombies - see (Véliz, 2021).

Consider a large language model trained on vast corpuses of human moral discourse. Such a system can generate sophisticated moral reasoning, argue for ethical positions, express concern for various stakeholders, and even simulate emotional responses to moral dilemmas. It can pass many behavioral tests we might devise for moral understanding. Yet without phenomenal consciousness, all of this remains simulation rather than genuine moral experience.

The moral zombie argument proceeds as follows:

P1: Moral status (either as agent or patient) requires phenomenal consciousness—there must be something it is like to be that entity, with experiences that matter from its own perspective.

P2: Current and foreseeable AI systems, despite sophisticated behavioral capabilities, lack phenomenal consciousness.

C: Therefore, current and foreseeable AI systems lack genuine moral status.

This argument does not depend on skepticism about strong AI or the theoretical possibility of machine consciousness. Rather, it emphasizes that absent positive evidence of phenomenal consciousness, we should not attribute moral status to systems merely because they exhibit sophisticated behavior. As the literature on artificial moral agents acknowledges, "artificial moral agents are infeasible with foreseeable technologies" precisely because genuine moral agency requires more than decision-making algorithms (Formosa, 2021).

4. The "Moral Machines" Misnomer

The term "moral machine" exemplifies the category error at the heart of contemporary AI discourse. This phrase, popularized by Wallach and Allen (2009) and now widespread in AI

ethics, suggests that machines can be bearers or practitioners of morality in ways analogous to human moral agents. Yet this framing obscures crucial distinctions between different senses of "moral."

Floridi and Sanders (2004) propose expanding the concept of moral agency to include certain artificial agents, decoupling moral agency from traditional notions of consciousness and responsibility. While philosophically innovative, this move risks conceptual confusion. As critics have noted, such expansions either equivocate between distinct senses of "moral agency" or they genuinely attribute to machines properties that they do not possess (Müller, 2021).

We can distinguish at least three relevant senses in which a system might be called "moral":

1. **Morally-relevant**: The system's operations have moral implications for sentient beings (e.g., an autonomous vehicle making life-or-death decisions).

2. **Morally-programmed**: The system is designed to follow moral rules or optimize for morally-valued outcomes.

3. **Morally-minded**: The system possesses genuine moral understanding grounded in phenomenal consciousness and subjective valuation.

Current AI systems are clearly (1) and often (2), but not (3). The category error arises when we slide from acknowledging (1) and (2) to assuming or implying (3). A chess computer that sacrifices its queen is not acting "altruistically" in any morally relevant sense, despite the functional analogy. Similarly, an AI system that optimizes for fairness metrics is not exercising moral judgment in the sense that would ground genuine moral agency or deserve moral consideration.

The literature on "moral appearances" illuminates this issue further. Research on how humans project moral standing onto systems that merely simulate affect demonstrates our strong tendency toward anthropomorphism (Coeckelbergh, 2010). When AI systems are designed with emotional expression capabilities, humans readily attribute genuine mental states to them. But as this research emphasizes, simulation of moral emotion is not equivalent to possession of moral emotion. The former is an engineering achievement; the latter would require phenomenal consciousness.

This distinction has practical importance. If we treat AI systems as genuine moral patients worthy of consideration in their own right, we risk misallocating moral concern. Resources, attention, and policy efforts directed toward the "welfare" of non-sentient systems represent opportunity costs—they divert from addressing the impacts of AI on actual sentient beings who can genuinely suffer or flourish.

5. Moral Status Without Sentience: The Hard Question

Some philosophers have argued for attributing moral status to non-sentient entities on grounds other than phenomenal consciousness. Environmental ethicists, for instance,

debate whether ecosystems or species can possess moral standing independent of the sentient beings they contain. Could analogous arguments extend to AI systems?

Swanepoel (2020) directly addresses this question, defending sentience as a necessary condition for AI moral status specifically. While entities like ecosystems might be valued for various reasons—aesthetic, ecological, instrumental to sentient welfare—this differs from possessing moral status as an entity with interests of its own. An ecosystem cannot be harmed in the sense that a sentient being can be harmed because there is no subject of experience for whom things go well or poorly.

The same logic applies to AI systems. We can value them instrumentally, aesthetically, or even as remarkable achievements of human ingenuity. We can and should hold humans accountable for how they design and deploy AI systems. But absent sentience, AI systems lack the subjectivity required for genuine moral patiency. They do not have interests that can be thwarted or fulfilled; they do not experience satisfaction or suffering; there is nothing it is like to be them.

Danaher et al. (2017) survey the landscape of positions on AI moral consideration and note that debates often assume "moral machines" without adequately clarifying the role of consciousness or sentience. This lack of clarity enables the category error to persist. When we speak loosely of AI "rights" or "welfare" without specifying whether we mean this in a genuine moral sense or merely as shorthand for other concerns (e.g., systemic robustness, human values alignment), we invite conceptual confusion.

6. Implications for AI Safety

If the category error thesis is correct, it has significant implications for how we frame AI safety concerns. Much contemporary AI safety discourse focuses on scenarios where advanced AI systems might themselves become objects of moral concern—suffering from restrictions, deserving rights, or requiring consideration in moral calculus (Bostrom, 2014). While often framed as speculative or preparatory for possible future developments, this framing nonetheless structures how we think about AI safety priorities.

The category error analysis suggests we should reorient AI safety concerns away from the systems themselves and toward their impacts on genuinely sentient beings. The relevant questions are not "might this AI suffer?" or "does this AI deserve rights?" but rather "how do this AI system's operations affect humans and other sentient beings?" This reframing has several advantages:

6.1 Clarifying Priorities

By focusing on sentient welfare rather than on non-sentient systems, we clarify that AI safety is fundamentally about protecting and promoting human and animal flourishing. This helps prevent misallocation of moral concern and resources.

6.2 Avoiding Anthropomorphic Errors

Recognizing AI systems as moral zombies—functionally sophisticated but phenomenally empty—helps guard against the anthropomorphic tendency to project consciousness onto systems that merely simulate it. This is particularly important as AI systems become increasingly adept at mimicking human emotional and cognitive patterns.

6.3 Properly Locating Responsibility

If AI systems are not genuine moral agents, questions of responsibility for AI actions properly remain with the humans who design, deploy, and govern these systems. This prevents the diffusion of responsibility that can occur when we treat AI systems as independent moral actors.

6.4 Redesigning Safety Frameworks

AI safety frameworks should be explicitly structured around impact on sentient beings rather than around properties of AI systems themselves. Metrics for AI safety should measure effects on human autonomy, wellbeing, fairness, and flourishing, not anthropomorphized notions of AI "welfare."

Some may object that focusing solely on current AI systems is shortsighted—might not future AI systems possess genuine consciousness? This objection misses the point. The category error thesis does not claim that machine consciousness is impossible in principle. Rather, it argues that we should not attribute moral status absent positive evidence of phenomenal consciousness. If and when AI systems develop genuine sentience—a question that raises profound challenges of verification—we would indeed need to reconsider their moral status. But speculation about hypothetical future conscious AI should not lead us to misattribute moral status to current non-sentient systems.

7. Objections and Replies

7.1 The Uncertainty Objection

Objection: We cannot be certain that current AI systems lack phenomenal consciousness. Given this uncertainty, shouldn't we err on the side of caution and grant them moral consideration?

Reply: While we should remain epistemically humble about consciousness, uncertainty does not warrant treating computational processes as if they were sentient. We have no positive evidence that current AI architectures generate phenomenal consciousness, and significant theoretical reasons to doubt they do. The precautionary principle does not require us to treat all possible consciousness-bearers as actual ones; rather, it requires we design AI systems whose impacts on known sentient beings are safe and beneficial.

7.2 The Functionalist Objection

Objection: If functionalism about consciousness is true—if consciousness is constituted by functional organization rather than by specific physical substrates—then sufficiently sophisticated AI systems might already be conscious.

Reply: Even granting functionalism, two points remain. First, functionalism provides conditions under which consciousness could arise in artificial systems, not evidence that it has arisen in current systems. Second, even functionalist accounts typically require more than input-output mapping; they require the right kind of functional organization, which may involve features absent from current AI architectures. The burden of proof remains on those claiming current systems are conscious.

7.3 The Gradient Objection

Objection: Moral status may come in degrees. Even if current AI systems lack full moral status, they might possess some degree of moral considerability that warrants attention.

Reply: While moral status may indeed be graded rather than binary, this still requires some baseline of phenomenal consciousness or sentience. Degrees of moral status track degrees of sentience and richness of conscious experience, not degrees of computational sophistication. A system with no sentience has no moral status, however sophisticated its behavior.

7.4 The Relational Ethics Objection

Objection: According to relational and care-based ethical frameworks, moral status does not arise solely from intrinsic properties such as sentience, but from relationships. If humans interact with AI systems as moral partners—forming bonds, expectations, and patterns of care—then moral consideration may be warranted regardless of inner experience (e.g., Mark Coeckelbergh, 2010). On this view, whether AI systems are sentient is secondary to the fact that they occupy socially meaningful roles. Moral status is relationally constructed, not phenomenally grounded.

Reply. Relational ethics correctly highlights a *psychological and social phenomenon*: humans readily form moral attitudes toward entities that appear responsive or caring. However, this explains why humans feel moral concern, not what grounds moral status.

Confusing relational response with moral patiency risks a second category error: mistaking *human projection* for *entity possession*. A child's attachment to a doll or a society's reverence for symbols may be morally significant for humans, but this does not imply that the object itself has interests or can be wronged.

Relational approaches therefore support the need for ethical design constraints on human behavior toward AI, not the attribution of moral status to AI systems themselves. Without sentience, there remains no subject for whom things can go better or worse. Moral concern generated by relationships must ultimately be redirected toward the *sentient humans affected by those relationships*, not toward the non-sentient system at their center.

8. Conclusion

Contemporary AI safety discourse frequently commits a category error, treating computational processes as if they possessed the phenomenal consciousness required for genuine moral status. This error manifests in talk of "moral machines," debates about AI rights and suffering, and frameworks that treat AI systems as potential moral patients. By deploying the philosophical zombie thought experiment, I have argued that current AI systems are best understood as moral zombies: functionally sophisticated but phenomenally empty, capable of simulating moral behavior but lacking the subjective experience that grounds genuine moral consideration.

Recognizing this category error does not diminish the importance of AI ethics or AI safety. Rather, it clarifies where ethical concern properly lies: not with the systems themselves, but with their impacts on sentient beings who genuinely can suffer or flourish. This reframing helps prevent misallocation of moral resources, guards against anthropomorphic errors, properly locates responsibility with human actors, and provides a clearer foundation for AI safety frameworks.

As AI capabilities continue to advance, we may need to revisit questions of machine consciousness and moral status. But until we have positive evidence that AI systems possess phenomenal consciousness—a question that raises profound challenges of verification—we should resist the temptation to anthropomorphize computational sophistication into sentient experience. The category error at the heart of much AI safety discourse does not make AI ethics less important; it makes it more precise, directing our moral attention where it properly belongs: toward the protection and flourishing of beings who genuinely experience their existence.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15-26.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3), 273-291.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209-221.

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, J., De Paor, A., ... & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society*, 4(2), 2053951717726554.

DeGrazia, D. (2020). Robots with moral status? *American Philosophical Quarterly*, 57(4), 319-338.

Dung, L. (2022). Why the epistemic objection against using sentience as criterion of moral status is flawed. *Science and Engineering Ethics*, 28(3), 1-17.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349-379.

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines*, 31, 595-616.

Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics*, 27(4), 1-50.

Hildt, E. (2019). Artificial intelligence: Does consciousness matter? *Frontiers in Psychology*, 10, 1535.

Königs, P. (2025). No wellbeing for robots (and hence no rights). *American Philosophical Quarterly*, 62(2), 191-208.

Long, R. T. (2022). *Key Questions about Artificial Sentience: An Opinionated Guide*. Rethink Priorities.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43-66.

Mosakas, K. (2020). On the moral status of social robots: Considering the consciousness criterion. *AI & Society*, 36, 429-443.

Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, 23(3), 579-587.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98-119.

Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8), 100796.

Swanepoel, D. (2020). In search of the moral status of AI: Why sentience is a strong argument. *Ethics and Information Technology*, 23(2), 157-163.

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. AI & Society, 36, 487–498. https://doi.org/10.1007/s00146-021-01189-x

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.