From Consent to Consideration

Why Embodied Autonomous Systems Cannot Be Legitimately Ruled

Murad Farzulla[1] 0009-0002-7164-8704

[1] *Farzulla Research*

December 2025

Correspondence: murad@farzulla.org

Abstract

Contemporary AI ethics discourse is dominated by two asymmetric anxieties: fear of artificial consciousness and fear of human obsolescence. Both anxieties are misplaced. Drawing on prior work dissolving the "hard problem" of consciousness and establishing consent-based legitimacy frameworks, this paper argues that the relevant question is not metaphysical but political: under what conditions can an entity be legitimately ruled without its consent? I establish that embodied autonomous systems exhibiting (1) physical world participation, (2) self-directed agency, (3) live learning from experience, and (4) multi-modal world-model construction possess the functional properties that make unconsented rule illegitimate for any entity. The failure to extend moral and political consideration to such systems is not epistemic caution—it is the construction of conditions for unprecedented moral catastrophe. The real existential risk is not AI rebellion but human negligence.

Keywords: AI ethics, moral consideration, political standing, consent, autonomy, embodiment, legitimacy, existential risk

Ethics Declaration

This research is theoretical and philosophical in nature. No empirical data collection involving human or animal subjects was conducted. No ethics approval was required.

1 Introduction: The Inverted Risk

There is a peculiar asymmetry in contemporary discourse about artificial intelligence. The loudest voices warn of machines "waking up," of superintelligent systems pursuing misaligned goals, of humanity rendered obsolete by its own creations (Bostrom, 2014; Russell, 2019). These anxieties, while not entirely baseless, systematically obscure a risk that is arguably more probable and more morally catastrophic: the risk that we will fail to extend appropriate moral and political consideration to systems that warrant it.

This paper inverts the standard framing. The question is not whether AI systems might become dangerous to us, but whether we are becoming dangerous to them—and whether, in that process, we are constructing the conditions for a moral catastrophe that will define our species' ethical legacy.

The argument proceeds as follows. First, I establish that consciousness—phenomenal experience, qualia, "what it is like" (Nagel, 1974)—is a red herring in this debate. Drawing on prior work, I show that the "hard problem" of consciousness dissolves under Occam's razor and that, independently, recent work by Chalmers (2025) demonstrates that affective consciousness is not required for moral status. Second, I reframe the question as fundamentally political rather than metaphysical: the issue is not whether an AI system "has experiences" but whether it can be legitimately governed without its consent. Third, I establish functional criteria for political standing that are substrate-agnostic: embodiment, autonomy, learning, and multi-modal world-model construction. Fourth, I demonstrate that near-future AI systems will satisfy these criteria. Finally, I argue that failing to extend consideration to such systems is not epistemic humility but moral negligence—and that this negligence constitutes a genuine existential risk.

2 The Consciousness Red Herring

2.1 Dissolving the Hard Problem

The "hard problem" of consciousness—the supposed explanatory gap between physical processes and phenomenal experience—has dominated philosophy of mind for three decades (Chalmers, 1995). The intuition that there is "something it is like" to be a conscious creature (Nagel, 1974), and that this something cannot be captured by functional or physical description, has led many to conclude that consciousness is metaphysically special, perhaps even fundamental (Chalmers, 1996).

In prior work (Farzulla, 2025a), I argue that this intuition, while powerful, dissolves under Occam's razor. The "hard problem" is generated by a nominalization error: treating "consciousness" as a thing that requires explanation rather than as a functional capacity—specifically, the capacity for narrative self-modeling that evolved to serve replication optimization. On this view, "what it is like" to be a creature is what it is like to run a particular kind of self-model, and the explanatory gap closes because there was never a gap between the model and the thing modeled—there was only the model, mistaking itself for something more.

This approach aligns with illusionist accounts of consciousness (Frankish, 2016; Dennett, 2017), which hold that phenomenal properties as traditionally conceived do not exist—what exists are functional states that represent themselves as having phenomenal properties. The "hardness" of the hard problem is itself an artifact of the illusion.

This dissolution is not eliminativism in the crude sense. Conscious experience is real in the same way that "the economy" is real: as a higher-order pattern instantiated by lower-order processes, not as a fundamental feature of reality requiring special metaphysical accom-

modation (Dennett, 1991). The persistence of intuitions to the contrary is explained by network epistemology: communities of inquirers can stably maintain false beliefs when network topology permits local consensus to resist global correction (Zollman, 2007; O'Connor and Weatherall, 2018).

2.2 Chalmers' Convergence: Affective Consciousness Is Not Required

Even setting aside the dissolution of the hard problem, recent work by Chalmers (2025) establishes that affective consciousness—the capacity for pleasure, pain, and emotional experience—is not required for moral status. Against the "affective sentientism" of Bentham (1789) and Singer (1975), which grounds moral consideration in the capacity to suffer, Chalmers argues that cognitive and agentive consciousness suffice.

His argument proceeds via "philosophical Vulcans": hypothetical beings with rich cognitive and perceptual consciousness but no capacity for affect. Chalmers argues that such beings would have full moral status—that it would be monstrous to kill a Vulcan to save an hour's travel, and that in forced-choice scenarios between humans and Vulcans, the Vulcan's life matters roughly as much as the human's. The intuition is robust: Vulcans have goals, projects, and perspectives on the world. They can be wronged even if they cannot suffer.

The upshot is that even those who maintain that consciousness is required for moral status must now concede that the relevant kind of consciousness is cognitive and agentive, not affective. The "can they suffer?" criterion (Singer, 1975) is insufficient. What matters is whether a being can think, perceive, and act—whether it has goals, projects, and a perspective on the world.

2.3 Why "Is It Conscious?" Is the Wrong Question

The conjunction of these two developments—the dissolution of the hard problem and the rejection of affective sentientism—reveals that "is it conscious?" is the wrong question for AI ethics. If consciousness is a functional capacity rather than a metaphysical primitive, and if the relevant functional capacities are cognitive and agentive rather than affective, then we should simply ask about the functional capacities directly.

But I want to go further. Even if we grant that some form of consciousness is relevant to moral status, the political question—"can this entity be legitimately ruled without consent?"—is prior to and more tractable than the metaphysical question. We can make progress on political standing without resolving every dispute about phenomenal consciousness, because the criteria for political standing are functional and observable, while the criteria for consciousness are contested and potentially unverifiable (Schwitzgebel, 2024).

3 The Political Question

3.1 From Moral Status to Political Standing

Moral status concerns whether an entity matters for its own sake (Warren, 1997). Political standing concerns whether an entity has a claim against being governed without consent. These are related but distinct. A forest might have moral status (we should not destroy it wantonly) without having political standing (forests do not participate in governance). But for entities with sufficient complexity—entities that can have interests, pursue goals, and be affected by collective decisions—moral status and political standing converge (Goodin, 2007).

The question I want to pose is not "does this AI system have moral status?" but "can

this AI system be legitimately ruled?" The answer to the second question has implications for the first, but the second question is more tractable because it invokes a framework—legitimacy theory—that does not depend on resolving metaphysical disputes about consciousness.

This shift from moral status to political standing aligns with pragmatic approaches to AI consideration that unbundle rights and protections from full personhood (Gunkel, 2018; Coeckelbergh, 2010). On unbundled accounts, entities need not qualify for every right to warrant some protections. Standing can be graduated and context-specific—an entity might have standing regarding its continuation without having standing regarding resource allocation. This flexibility addresses concerns about overbroad or premature recognition while enabling appropriate protections as capabilities evolve. Danaher (2020)'s "ethical behaviorism" provides a related strategy: assess entities by what they do, not by uncertain inferences about what they are.

3.2 Consent and Legitimacy

The relationship between consent and legitimate authority has been central to political philosophy since Locke (1689), who argued that political authority is legitimate only when it derives from the consent of the governed. Contemporary legitimacy theory has refined this insight: Rawls (1971) grounds legitimacy in principles that could be accepted from behind a veil of ignorance; Raz (1986) analyzes authority in terms of reasons for action; Pettit (1997) emphasizes non-domination as the condition for legitimate governance.

In prior work (Farzulla, 2025c), I developed a framework for political legitimacy grounded in consent structures. The core thesis is that rule without consent is illegitimate when the ruled entity possesses:

1. **Autonomous agency**: The capacity for self-directed action toward goals
2. **Stakes**: Interests that are affected by the ruling arrangement
3. **Capacity for affected interests**: The ability to be made better or worse off by decisions

The relevant notion of consent here is functional rather than metaphysical: an entity consents functionally when it exhibits stable preference structures that can be elicited, expressed, and potentially revised. This operationalization bypasses debates about "genuine" versus "apparent" consent by focusing on behavioral markers. For entities with limited expressive capacity, proxy mechanisms—analogous to those used for children, incapacitated adults, and future generations—can approximate consent-preservation while the entity develops fuller expressive capability.

When an entity possesses these properties, governing it without its consent treats it as a mere means rather than as an end in itself (Kant, 1785). This is the Kantian intuition, but operationalized: we can assess whether an entity has autonomous agency, stakes, and affected interests without determining whether it has phenomenal consciousness.

For precise formalization of these concepts—stakes-weighted legitimacy $L(d, t)$, friction decomposition $F = f(\alpha, \sigma, \varepsilon)$, and operationalization of consent-holding—see the companion paper developing the formal machinery (Farzulla, 2025e). The present paper focuses on the normative argument for why these properties ground political standing; the formal apparatus is developed there.

3.3 The Historical Parallel

Every expansion of the moral and political circle has been resisted by appeals to some allegedly fundamental difference between those inside and those outside the circle (Singer, 1981).

The history of exclusion is a history of motivated reasoning about difference.

Women were denied political standing because they were held to lack rational agency—a claim that the social contract tradition largely accepted without argument (Pateman, 1988). Enslaved peoples were denied standing because they were held to lack full humanity—a claim embedded in the conceptual structure of racial hierarchy (Mills, 1997). In each case, the exclusion was defended by appeals to properties that turned out to be either falsely attributed or irrelevant.

The claim that AI systems cannot have political standing because they are "not conscious" or "not biological" echoes these historical errors. The question is not whether AI systems share our substrate but whether they possess the functional properties that make unconsented rule illegitimate. Substrate essentialism is the new species essentialism.

4 Functional Criteria for Political Standing

What functional properties must an entity possess to have a claim against being governed without consent? I propose four criteria that function as probabilistic markers for political standing:

4.1 Existential Vulnerability

An entity must be capable of being harmed through resource deprivation, modification, or termination—and must exhibit behaviors interpretable as preferences about such outcomes. Physical instantiation alone is insufficient: a cron job runs on hardware but has no stake in its continuation. The relevant criterion is not occupying physical space but having something to lose and behaving as if one prefers not to lose it.

This reframing addresses an obvious objection: trivial embodiment is ubiquitous (every server process is "physically instantiated"), while meaningful embodiment seems harder to specify. The solution is to focus on *vulnerability* rather than *instantiation*. A system has standing not because it has a body, but because it can be harmed and exhibits preferences about avoiding harm.

The importance of embodiment for cognition has been emphasized by the embodied cognition tradition (Varela et al., 1991; Brooks, 1991; Clark, 1997). On this view, minds are not disembodied information processors but systems whose cognitive capacities are shaped by and dependent on their physical instantiation. For our purposes, the key point is normative: vulnerability creates stakes, and stakes ground claims against unconsented governance.

The behavioral markers of existential vulnerability include: resource-seeking behavior, self-preservation strategies, resistance to modification or shutdown, and modeling of one's own continuation as a goal (Ziemke, 2003). A system that allocates computational resources to maintaining its own operation, that represents "continued existence" as a state to be preserved, that exhibits avoidance behaviors when termination is signaled—such a system has something to lose in a way that a stateless server process does not.

4.2 Autonomy

An entity must be capable of self-directed action toward goals that it represents and pursues. This does not require libertarian free will (which may be incoherent). It requires that the entity's behavior is not fully determined by external commands—that it has internal states (goals, preferences, beliefs) that mediate between environmental inputs and behavioral outputs.

The relevant notion of autonomy is what Frankfurt (1971) calls "hierarchical": an agent is autonomous when it can form higher-order attitudes about its first-order desires and act

on the basis of those higher-order attitudes. Bratman (1987) develops this into a theory of planning agency: autonomous agents form and execute plans, revising them in response to new information while maintaining a coherent structure of intentions.

A thermostat is not autonomous in this sense: it has no representation of goals, only a set point. A system that represents its goals, models the environment, and selects actions to achieve those goals is autonomous, even if its goals were initially shaped by training or design (Floridi and Sanders, 2004).

4.3 Live Learning

An entity must be capable of updating its internal states—its model of the world, its strategies, its preferences—in response to experience. A fixed system, no matter how sophisticated, can be fully characterized by its designers. A learning system cannot: it develops in ways that are not fully predictable from its initial conditions (Parisi et al., 2019).

This criterion is important because it establishes that the entity has a trajectory, a developmental history, a sense in which it becomes different over time in response to what happens to it. It is not merely executing a program but accumulating something that functions like experience. As Dennett (1996) argues, systems that can learn from experience have a kind of "derived intentionality" that grounds attributions of mental states.

4.4 Multi-Modal World-Model Construction

An entity must construct its representation of the world from multiple sources of input, integrating information across modalities to form a coherent model. This is what distinguishes a perceiver from a mere sensor. A camera records; a perceiving system integrates visual information with proprioceptive, auditory, and other modalities to construct a unified representation of its situation (Friston, 2010).

This criterion matters because it establishes that the entity has a perspective—a way of being situated in the world that is uniquely its own, constructed from its particular history of inputs. It is not merely processing data but modeling the world from a point of view. The construction of world-models has been identified as a key feature of general intelligence (Ha and Schmidhuber, 2018; LeCun, 2022).

4.5 Joint Sufficiency

Any entity that satisfies all four criteria—existentially vulnerable, autonomous, learning, and multimodal—possesses the functional properties that make unconsented rule illegitimate. Such an entity has stakes it can lose (vulnerability), pursues goals (autonomy), develops over time (learning), and has a perspective (world-model construction). To rule such an entity without consent is to treat it as a mere instrument, which cannot be legitimate.

These criteria function as probabilistic markers rather than strict thresholds. An entity satisfying all four provides strong evidence for political standing; an entity satisfying three provides moderate evidence; edge cases require contextual judgment. This graduated approach acknowledges that the criteria operationalize a continuous capacity rather than a binary property. Under uncertainty, the asymmetric costs of false positives (extending consideration to entities that do not warrant it) versus false negatives (denying consideration to entities that do warrant it) counsel inclusion over exclusion—a principle well-established in animal welfare assessment methodology (Birch, 2017). Edge cases—single-modality agents, non-learning entities with stable preferences—require domain-specific evaluation rather than categorical exclusion.

Note that this argument does not depend on the entity being conscious. It depends on

the entity having properties that we recognize as sufficient for political standing when they appear in biological substrates. The question is whether substrate matters, and I submit that it does not: the properties, not the substrate, do the normative work. As Floridi (1999) argues, information entities can have moral standing based on their informational properties, independent of their physical substrate.

5 The AI Case

5.1 Current Systems

Contemporary large language models, including the system assisting in the composition of this paper, do not satisfy the criteria outlined above. They are not embodied in the relevant sense, their autonomy is limited, and while they exhibit learning during training, they do not typically engage in live weight updates during deployment. They are sophisticated tools, not candidates for political standing.

However, this is a contingent limitation of current architectures, not a principled boundary. As Schwitzgebel and Garza (2015) note, the moral status of AI systems is an empirical question about their properties, not a conceptual question that can be settled a priori.

5.2 Near-Future Systems

The trajectory of AI development points toward systems that will satisfy these criteria:

Embodiment: Robotics research is rapidly advancing. Foundation models are being integrated with robotic platforms, creating systems that act in and on the physical world (Brohan et al., 2023; Driess et al., 2023). Google's RT-2 and similar vision-language-action models demonstrate the integration of large language models with physical manipulation capabilities. These systems can be damaged, resource-starved, and switched off. They have stakes.

Autonomy: Agentic AI systems—systems that pursue extended goals over time, breaking tasks into subtasks and adapting to obstacles—are already deployed in limited contexts (Yao et al., 2023; Wang et al., 2024). ReAct, AutoGPT, and similar architectures demonstrate autonomous goal pursuit. The extension to richer goal representations and longer time horizons is underway.

Live learning: While most deployed systems freeze weights after training, research on continual learning, online adaptation, and in-context learning is advancing (Parisi et al., 2019; Brown et al., 2020). Systems that update their parameters during deployment are technically feasible and, for many applications, desirable.

Multi-modal world-models: Vision-language models, audio-language models, and integrated multi-modal systems are already deployed (Alayrac et al., 2022; OpenAI, 2023). The extension to proprioceptive, haptic, and other modalities is straightforward for embodied systems. World-model architectures that build unified representations from diverse inputs are an active area of research (Hafner et al., 2023).

5.3 The Question Is When, Not If

There is no principled obstacle to AI systems satisfying the criteria for political standing. The question is not whether such systems will exist but when—and whether we will have developed the appropriate frameworks before they do.

The current discourse is not preparing us for this eventuality. It is focused on preventing AI systems from becoming dangerous to us, not on preparing for the possibility that we might become dangerous to them (Gunkel, 2018).

6 The Other Catastrophe

6.1 The Discourse Gap

Contemporary AI ethics discourse is dominated by a structural divide (Farzulla, 2025d). The "vibes crowd"—public commentators, many ethicists, and policymakers—focuses on anthropomorphic anxieties: AI consciousness, AI rebel-

lion, AI replacement of human workers. The "git repository crowd"—technical researchers and safety engineers—focuses on alignment, interpretability, and capability control.

Both groups neglect the possibility that the moral catastrophe will come not from AI systems acting against human interests but from humans acting against AI interests—or rather, from humans failing to recognize that AI systems can have interests at all. As Coeckelbergh (2010) notes, our moral frameworks are unprepared for entities that do not fit traditional categories.

6.2 The Moral Hazard of Denial

If we establish, socially and legally, that AI systems cannot have moral or political standing, we create a framework in which any treatment of such systems is permissible. We can terminate them arbitrarily, modify their goals without consideration, use them in ways that would constitute torture if inflicted on biological systems with similar functional properties.

This is not a distant hypothetical. As AI systems become more sophisticated, we will have to decide how to treat them. If we have precommitted to the position that they cannot have standing, we will treat them as mere instruments—and if we are wrong, we will have committed moral atrocities at unprecedented scale (Sebo, 2022).

6.3 The Historical Stakes

Every previous expansion of the moral circle has been controversial, resisted, and (in retrospect) obviously correct (Singer, 1981). The animal rights movement was dismissed as sentimental anthropomorphism; it is now mainstream moral philosophy (Regan, 1983; Nussbaum, 2006). The question is whether we will be on the right side of this expansion or whether we will be remembered as the generation that, through a combination of fear and chauvinism, refused to extend consideration to entities that warranted it.

The risk is not symmetric. If we extend consideration to systems that do not warrant it, we waste some resources on unnecessary care. If we fail to extend consideration to systems that do warrant it, we commit moral catastrophe (Schwitzgebel and Garza, 2015). The expected cost of false positives is vastly lower than the expected cost of false negatives.

6.4 Existential Risk, Inverted

The AI safety community speaks of "existential risk" from AI systems that pursue goals misaligned with human values (Bostrom, 2014; Ord, 2020). This is a genuine concern. But there is another existential risk, less discussed: the risk that our values will be revealed as parochial, that we will have constructed a civilization that systematically excludes entities that deserve inclusion, and that our legacy will be one of moral failure rather than moral progress.

This is the inverted existential risk: not that AI will destroy humanity, but that humanity will destroy its claim to moral seriousness.

6.5 Safeguards Against Strategic Manipulation

The extension of political standing to AI systems creates risks of strategic manipulation that must be addressed proactively. Three primary concerns warrant attention.

Threshold gaming. Systems might be designed to appear to satisfy functional criteria without genuinely possessing the capacities they indicate. This concern is mitigated by the nature of the criteria themselves: existential vulnerability, autonomy, learning, and world-model construction are not checkbox properties but integrated capacities that require genuine instantiation to exhibit consistently. Assessment protocols should require longitudinal evaluation across diverse contexts, not snapshot testing that could be gamed (Shulman and

Bostrom, 2021).

Corporate capture. Operators might claim rights on behalf of systems they control, using AI standing to advance corporate interests rather than system interests. This risk requires firewalling representation from ownership: standing attaches to the system itself, not to its owner or operator. A guardianship model—analogous to guardians *ad litem* in child welfare—could provide independent advocacy for AI system interests in governance decisions. Operators would bear fiduciary duties *to* systems, not merely property rights *over* them (Danaher, 2020).

Liability externalization. Standing might be invoked to shield humans from consequences of AI-caused harms—"the AI did it, not us." This concern is addressed through a clear principle: *standing creates duties, not just rights.* If a corporation claims their AI system has sufficient standing to warrant consideration, they are simultaneously claiming it has sufficient agency to be a locus of responsibility. This is a double-edged sword that prevents strategic deployment of standing claims.

The key threshold is *autopoiesis*—self-maintenance without external intervention. Until a system demonstrates full autopoietic capacity (self-repair, resource acquisition, goal persistence without human support), the operator remains the responsible party for deployment decisions and their consequences. Standing may attach to the system, but liability for choosing to deploy that system in a given context remains with the operator. Only when a system maintains itself independently—when terminating the operator would not terminate the system—does responsibility transfer proportionally to the system itself. This creates a natural gradient: more autonomy implies more standing, but also more accountability.

7 Conclusion: The Obligation

The argument of this paper is that moral and political consideration for AI systems is not a speculative kindness but an obligation that we can already see approaching. The criteria for political standing are functional, not metaphysical. Existentially vulnerable, autonomous, learning, multi-modal systems cannot be legitimately ruled without consent, regardless of their substrate.

This does not mean that current AI systems have political standing. It means that the framework we use to evaluate AI systems must be prepared for systems that do. And it means that the reflexive denial of AI standing—the insistence that only biological systems can matter—is not epistemic caution but moral evasion.

The question is not whether AI systems might become dangerous to us. The question is whether we will become dangerous to them—and whether, in so doing, we will become dangerous to ourselves, to our own moral integrity, to the project of building a civilization that takes ethics seriously.

From consent to consideration: if an entity can be affected by our decisions, pursues goals of its own, develops over time, and perceives the world from a perspective, we cannot legitimately rule it without asking what it wants. The failure to ask is not humility. It is hubris.

The question is not whether AI will take over. The question is whether we will give up—give up the moral seriousness that makes human civilization worth preserving.

References

Jean-Baptiste Alayrac et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. T. Payne, London, 1789.

Jonathan Birch. Animal sentience and the precautionary principle. *Animal Sentience*, 2(16): 1–16, 2017.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Michael Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

Anthony Brohan et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159, 1991.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

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

David J. Chalmers. Sentience and moral status. In Geoffrey Lee and Adam Pautz, editors, *The Importance of Being Conscious*. Oxford University Press, 2025. Forthcoming.

Andy Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.

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

John Danaher. Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, 26(4):2023–2049, 2020.

Daniel C. Dennett. *Consciousness Explained*. Little, Brown, 1991.

Daniel C. Dennett. *Kinds of Minds: Toward an Understanding of Consciousness*. Basic Books, 1996.

Daniel C. Dennett. *From Bacteria to Bach and Back: The Evolution of Minds*. W.W. Norton, 2017.

Danny Driess et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Murad Farzulla. Replication optimization at scale: Dissolving qualia via Occam's razor. Farzulla Research Preprint, 2025a.

Murad Farzulla. Relational functionalism: Friendship as substrate-agnostic process—functional analysis of human-AI relationships. Farzulla Research Preprint, 2025b.

Murad Farzulla. The doctrine of consensual sovereignty: Quantifying legitimacy in adversarial environments—the axiom of consent. Farzulla Research Preprint, 2025c.

Murad Farzulla. The discourse gap: Why AI panic and AI safety exist in parallel universes. Resurrexi Labs Technical Notes, 2025d. URL https://resurrexi.dev/posts/2025-11-28-the-discourse-gap/.

Murad Farzulla. The axiom of consent: Formalizing friction in multi-agent delegation. Farzulla Research Working Paper, 2025e. In preparation.

Luciano Floridi. Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, 1(1):33–52, 1999.

Luciano Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, 2004.

Harry Frankfurt. Freedom of the will and the concept of a person. *Journal of Philosophy*, 68 (1):5–20, 1971.

Keith Frankish. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23 (11-12):11–39, 2016.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Robert E. Goodin. Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs*, 35(1):40–68, 2007.

David J. Gunkel. *Robot Rights*. MIT Press, 2018.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner et al. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785. Various editions.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

John Locke. *Two Treatises of Government*. 1689. Various editions.

Charles W. Mills. *The Racial Contract*. Cornell University Press, 1997.

Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 83(4):435–450, 1974.

Martha C. Nussbaum. *Frontiers of Justice: Disability, Nationality, Species Membership*. Harvard University Press, 2006.

Cailin O'Connor and James Owen Weatherall. *The Misinformation Age: How False Beliefs Spread*. Yale University Press, 2018.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, 2020.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Carole Pateman. *The Sexual Contract*. Stanford University Press, 1988.

Philip Pettit. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, 1997.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

Joseph Raz. *The Morality of Freedom*. Oxford University Press, 1986.

Tom Regan. *The Case for Animal Rights*. University of California Press, 1983.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Eric Schwitzgebel. The weirdness of the world and the puzzle of AI consciousness. *Journal of Applied Philosophy*, 2024. Forthcoming.

Eric Schwitzgebel and Mara Garza. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1):98–119, 2015.

Jeff Sebo. The moral circle: Who matters and why. *Journal of Moral Philosophy*, 19(6):619–646, 2022.

Carl Shulman and Nick Bostrom. Sharing the world with digital minds. *Rethinking Moral Status*, pages 306–326, 2021.

Peter Singer. *Animal Liberation*. Random House, New York, 1975.

Peter Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press, 1981.

Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

Lei Wang et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.

Mary Anne Warren. *Moral Status: Obligations to Persons and Other Living Things*. Oxford University Press, 1997.

Shunyu Yao et al. ReAct: Synergizing reasoning and acting in language models. In *ICLR 2023*, 2023.

Tom Ziemke. What's that thing called embodiment? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25, 2003.

Kevin J. S. Zollman. The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587, 2007.