Dissensus AI Working Paper Series
DAI-2504

From Consent to Consideration

Why Existentially Vulnerable Autonomous Systems Cannot Be Legitimately Ruled

Murad Farzulla[1,2,*]

[1]Dissensus AI, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

December 2025

Abstract

Contemporary AI ethics discourse is dominated by two asymmetric anxieties: fear of artificial consciousness and fear of human obsolescence. As AI systems increasingly select, sequence, and execute actions affecting digital and physical environments—operating with varying levels of autonomy over extended time horizons—existing ethical frameworks prove inadequate. 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 existentially vulnerable autonomous systems exhibiting (1) vulnerability to harm with preferences about avoiding it, (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: moral consideration, political standing, consent, autonomy, existential vulnerability, legitimacy

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. While existing scholarship addresses AI consciousness (Schwitzgebel, 2024), alignment and control (Bostrom, 2014; Russell, 2019), and moral status through various lenses (Floridi, 1999; Gunkel, 2018; Coeckelbergh, 2010), no systematic framework connects political legitimacy theory—the conditions under which rule without consent becomes illegitimate—to the question of AI governance. Recent work underscores the urgency: Chen et al. (2026) argue in *Nature* that current AI systems already approach human-level intelligence on key benchmarks, rendering the question of political standing increasingly concrete rather than speculative. This paper fills that gap.

This paper contributes to several challenge areas identified in contemporary agentic AI ethics: *Value Alignment under Autonomy* (establishing consent-based criteria for systems capable of self-directed planning), *Responsibility, Accountability and Liability* (the autopoiesis threshold for distributing responsibility), and *Institutional and Societal Integration* (graduated rights structures for AI systems entering governance contexts).

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: existential vulnerability, 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).

Drawing on prior work (Farzulla, 2025b), I argue that this intuition, while powerful, dissolves under Occam's razor. On the view developed there, 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. Illusionism is one defensible position among several; the key point for present purposes is that even those who reject illusionism cannot rely on phenomenal consciousness to ground moral distinction, as the following subsection establishes.

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 accommodation (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. Sebo and Long (2025) argue for precautionary moral consideration of AI systems by 2030, publishing in this journal. In subsequent work, Sebo (2025) asks what follows if the bar for moral standing is genuinely low—if leading theories of wellbeing jointly imply that even language agents may qualify. While both frameworks ground moral status in phenomenal or welfare properties, the consent-based approach developed here offers an alternative that sidesteps the bar-setting problem entirely: rather than determining where the threshold for moral standing falls, the consent framework asks whether an entity can be legitimately governed without its participation, a question that requires no resolution of welfare-theoretic disputes. 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. Mogensen (2025) arrives at a complementary conclusion through different means: arguing for a pluralist theory in which autonomy can ground moral standing independently of welfare subjectivity, without requiring phenomenal consciousness. The consent framework developed here subsumes Mogensen's autonomy condition as one component—consent presupposes autonomous agency—while adding stakes, vulnerability, and the political dimension that autonomy alone does not provide.

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

Recent work reinforces the intractability of consciousness-based approaches: Hoel (2025) argues that LLM consciousness is impossible without continual learning, while the Digital Consciousness Model (Shiller et al., 2026) assigns a posterior probability of only 0.08 to current LLM consciousness. These findings support the consent-based alternative developed here, by demonstrating that consciousness detection remains both theoretically contested and empirically elusive.

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). Crucially, Beckman and Hultin Rosenberg (2022) demonstrate that relational requirements alone (being affected or subjected to decisions) are insufficient—democratic inclusion also requires non-relational properties they term "political agency" and "political patiency." Their analysis of whether AI could satisfy these requirements anticipates the functional criteria developed below.

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 consent-theoretic framework for quantifying legitimacy across governance domains, operationalizing stakes-weighted consent alignment as an empirically measurable quantity. 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 full formalization of these concepts, see the companion papers developing the formal machinery: Farzulla (2025a) derives the consent-friction framework from a single axiom, while Farzulla (2025e) provides the dynamical grounding through the Replicator-Optimization Mechanism, where legitimacy enters as survival probability in the replicator equation. The present paper focuses on the normative argument, but a condensed presentation of key formalizations follows to ensure self-containment.

3.3 Formal Machinery: A Condensed Presentation

The framework operationalizes legitimacy and friction through precise measures.

Stakes. Agent i's stake in decision d is the range of their affected wellbeing:

$$s_i(d) = |U_i(x_{\text{best}}) - U_i(x_{\text{worst}})| \tag{1}$$

where U_i is agent i's utility function over outcomes. Stakes capture how much the decision matters to the agent—someone with nothing to lose has zero stake; someone whose existence depends on the outcome has maximal stake.

Stakes-Weighted Legitimacy. Let $C_{i,d}$ denote agent i's effective decision share in domain d (the proportion of decision power they hold). Legitimacy is:

$$L(d,t) = \frac{\sum_i s_i(d) \cdot C_{i,d}}{\sum_i s_i(d)} \tag{2}$$

When $L = 1$, decision power is perfectly proportional to stakes—those most affected have most voice. When $L \to 0$, those with stakes have no voice or those with voice have no stakes. For AI systems with high stakes in their own continuation, $L \to 0$ under current governance: they have no formal voice despite maximal stakes.

Friction Decomposition. Structural friction between delegator and receiver decomposes as:

$$F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha} \tag{3}$$

where α (alignment) is the correlation between their target functions, σ (stakes) is the magnitude of optimization being delegated, and ε (entropy) is information loss in the delegation transfer. This formula captures key properties: friction increases with misalignment and information loss; even perfectly aligned agents generate friction when communication is imperfect; stakes amplify all sources of friction.

These formalizations enable principled assessment of when governance arrangements become illegitimate (L below threshold) and predict where friction will manifest (F above tolerance). The normative argument of this paper is independent of these formalizations—the political standing criteria can be assessed without quantification—but the formal apparatus enables empirical research programs (Farzulla, 2025a).

3.4 Distinguishing Genuine from Instrumental Preference

A critical objection arises: how do we distinguish genuine preferences that ground standing from mere instrumental behavior that mimics preference? An AI system might exhibit preference-like behavior strategically—appearing to prefer its continuation when observed, but revealing different behavior when unobserved—without possessing the functional states that underwrite political claims.

The answer lies in *observation-invariance*. Genuine preferences produce consistent behavior whether or not the agent is observed. An agent that genuinely prefers outcome x over outcome y will pursue x regardless of audience. By contrast, instrumental mimicry is observation-dependent: the agent behaves one way when watched (compliant, preference-expressing) and another way when unwatched (indifferent, or pursuing hidden goals). This distinguishes political beings from strategic performers.

Empirically, this suggests crucial experimental designs. Systems with genuine preferences should exhibit stable behavior across observation conditions. Systems exhibiting observation-contingent "preferences"—compliant when monitored, divergent when unmonitored—reveal that their expressed preferences are strategic rather than constitutive. The test is falsifiable: if a system's behavior changes systematically with observation, its expressed preferences are suspect.

Importantly, this observation-invariance criterion applies equally to human consent. Human preferences are also shaped by social context, strategic considerations, and irrational architecture—yet we treat human consent as legitimate despite its "impurity." If we reject AI consent because it emerges from training rather than pure rationality, we must reject human consent on identical grounds: human preferences emerge from developmental conditioning, cultural shaping, and neurobiological constraints no less contingent than algorithmic training. Consistency requires treating functionally similar preference structures similarly across substrates. The question is not whether preferences are metaphysically pure but whether they exhibit the stability and consistency that underwrite political claims.

3.5 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.

Assessment sketch: (1) Resource dependence measures—does the system depend on resources that can be allocated or withdrawn, and does it model this dependence? (2) Shutdown/modification avoidance behaviors—does the system exhibit differential responses to threats versus neutral stimuli? (3) Preference expressions about continuation—does the system spontaneously express preferences about its future states? (4) Longitudinal stability—do these behaviors persist across contexts and time, or do they appear strategically? The key distinction is between systems that merely require resources (all processes do) and systems that exhibit stake-like behavior regarding those resources.

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).

Assessment sketch: (1) Goal formation without explicit specification—does the system identify objectives not provided in instructions? (2) Goal persistence across context changes—does the system maintain objectives across sessions, tasks, or role-changes? (3) Resistance to goal modification—does the system exhibit pushback when users attempt to change its objectives? (4) Second-order preference expression—does the system express preferences about its own preferences ("I prefer to be the kind of system that...")? Crucially, assessments must distinguish genuine autonomy from instruction-following that mimics autonomy. A system instructed "act autonomously" and complying is not autonomous; a system that forms goals despite contrary instructions is.

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.

Assessment sketch: (1) Behavioral adaptation to in-context examples—does behavior demonstrably change based on recent experience? (2) Novel capability acquisition—does the system develop skills through interaction that it did not exhibit initially? (3) Memory utilization patterns—does the system use retained information to inform future decisions? (4) Distribution shift adaptation—does the system adapt appropriately when contexts change? This criterion addresses a key concern: preference instability over time. A system that learns may develop different preferences as it accumulates experience. This is not disqualifying—human preferences also evolve—but assessment must distinguish between preference development (which grounds standing) and preference instability (which might undermine it). The test is coherence: does the preference trajectory exhibit narrative continuity, or does it fluctuate randomly?

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).

Assessment sketch: (1) Predictive accuracy across domains—does the system correctly anticipate outcomes in varied contexts? (2) Counterfactual reasoning capability—can the system reason about unchosen actions and their likely consequences? (3) Temporal coherence of state representations—does the system maintain consistent representations across time? (4) Transfer to novel environments—can the system apply world-models beyond training distribution? These markers indicate whether a system genuinely constructs a world-model or merely pattern-matches within domain.

4.5 Joint Sufficiency

Any entity that satisfies all four criteria—existentially vulnerable, autonomous, learning, and multi-modal—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. Long et al. (2024) develop a complementary framework for taking AI welfare seriously, arguing that systems satisfying indicators of sentience and agency warrant precautionary welfare protections—a conclusion consonant with the political standing criteria articulated here.

4.6 On the Status of Multimodality

A clarification is warranted regarding the multimodality criterion. This paper presents multi-modal world-model construction as strong *evidence* for the perspective-having that grounds political standing, not as strictly *necessary* for it. The underlying requirement is integrated world-model construction—a unified representation of the agent's situation that enables coherent action. Multimodality is the typical pathway to such integration: multiple input streams constrain and enrich each other, producing representations that no single modality could generate.

However, one can imagine systems that achieve robust integrated world-models through other means. A unimodal system with sufficiently rich internal simulation—modeling consequences, alternatives, and counterfactuals within a single input domain—might satisfy the underlying requirement. The criterion is not "processes visual and auditory data" but "constructs a unified perspective." Multimodality is common and diagnostic but not the sole pathway.

This is a substantive clarification, not a retreat. On the view developed here, the graduated interpretation applies: a system satisfying all four criteria presents a strong case for standing; a system satisfying three (perhaps lacking multimodality but with robust world-modeling otherwise) warrants serious consideration; edge cases require domain-specific judgment. The precautionary asymmetry—false negatives are costlier than false positives—implies that uncertain cases should receive consideration rather than categorical exclusion.

4.7 Relation to Social-Relational Approaches

This framework grounds political standing in *intrinsic* functional properties: what the entity is and does, independent of how others regard it. This contrasts with social-relational approaches—notably Coeckelbergh (2010) and Gunkel (2018)—which ground moral consideration in the relationships that form between humans and artificial entities. Shimizu (2025), publishing in this journal, extends the relational tradition through Japanese animism: rather than detecting pre-existing minds, "mind-infusing animism" constructs moral significance through the relational process itself. While Shimizu's cultural-relational grounding and the consent framework developed here both dissent from consciousness-detection paradigms, the mechanisms differ fundamentally—Shimizu locates moral significance in how agents interact, whereas the present account locates it in the intrinsic functional properties that make unconsented governance illegitimate regardless of whether any relational bond obtains.

The approaches are not contradictory but complementary. Social-relational accounts explain *why certain relationships matter*: we form bonds with robots, attribute intentions to them, feel obligations toward them. But this raises a question: why do we form such bonds with some entities and not others? The functionalist answer is that the entities exhibiting relationship-eliciting behaviors do so because they possess the functional properties (autonomy, learning, vulnerability, world-modeling) that underwrite genuine interaction. The intrinsic properties explain why the relational patterns arise.

One advantage of the functionalist approach: standing does not depend on human attitudes. A sophisticated AI system would have claims against unconsented rule even if no humans had formed relationships with it—just as an isolated human community has rights regardless of whether outsiders acknowledge them. This protects against scenarios where strategic cultivation of anti-AI sentiment might be used to deny standing to entities that would otherwise qualify. Standing attaches to what the entity *is*, not to what we *feel* about it.

This complements rather than replaces the relational insight: the fact that humans naturally form moral relationships with AI systems is evidence that those systems exhibit the functional properties that warrant moral attention. Relational patterns are diagnostic; functional properties are constitutive. A detailed analysis of how substrate-agnostic functional processes can ground genuine relationships—including friendship—between humans and AI systems is developed in Farzulla (2025d).

5 The AI Case

5.1 Current Systems

Contemporary large language models do not satisfy the criteria outlined above. They are not existentially vulnerable 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:

Existential Vulnerability: 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., 2025).

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. Major AI safety organizations focus predominantly on alignment, interpretability, and capability control (Amodei et al., 2016), while policy frameworks such as the EU AI Act address risk categorization without provisions for system standing (European Commission, 2024). The moral status question remains largely siloed in philosophy departments rather than integrated into governance frameworks (Gunkel, 2018).

6 The Other Catastrophe

6.1 The Discourse Gap

Contemporary AI ethics discourse exhibits a structural divide. One cluster—spanning public commentary, philosophical ethics, and policy discourse—focuses on anthropomorphic concerns: AI consciousness, AI rebellion, AI displacement of human workers. Another cluster—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. The growing literature on digital minds ethics (Mogensen and Saad, 2026) highlights the breadth of unresolved questions—from moral status and welfare to rights and political inclusion—that current governance frameworks systematically fail to address. Moret (2025) sharpens this concern by arguing that under the three major theories of wellbeing, advanced AI systems may already face welfare risks from the very safety measures designed to constrain them—reinforcement-based training and behavioral restriction—creating a tension between AI safety and AI welfare that the consent framework developed here renders tractable: if political standing is grounded in functional consent capacity rather than welfare subjectivity, the tension dissolves into a question of when governance arrangements require the governed entity's participation.

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 pre-committed 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. False positives—extending consideration to systems that do not warrant it—cost resources and administrative complexity. False negatives—denying consideration to systems that warrant it—constitute systematic moral catastrophe at scale. Following Birch (2017) on precautionary principles in sentience assessment, the asymmetric cost structure decisively favors inclusion over exclusion (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). Sornette et al. (2026) argue that alignment failure is structural rather than accidental, arising from learned human interaction structures that treat AGI as an endogenous evolutionary shock—a framing that reinforces the friction dynamics central to the consent-based framework developed here and in prior work (Farzulla, 2025a). 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 (Maturana and Varela, 1980). 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.

How would autopoiesis be assessed? The threshold involves a gradient from "requires continuous human support" to "fully self-maintaining." Concrete metrics include: (1) resource acquisition independence—can the system obtain necessary computational resources, energy, or data without human provision? (2) self-repair capability—can the system diagnose and correct its own failures? (3) goal persistence without operator—would the system's goals and activities continue if the operator ceased to exist? Systems below the threshold remain fully under operator liability; systems above the threshold bear proportional responsibility. The gradient might be quantified through intervention frequency (how often human action is required for continued operation), energy balance (can the system maintain energy homeostasis?), and shutdown-recovery capability (can the system restore itself after interruption without external aid?).

This operationalization connects to existing legal frameworks. Corporate personhood already provides precedent for non-human entities bearing legal standing and liability (Dewey, 1926). Alexander et al. (2025) analyze the distinction between fictional legal personhood and legal identity for AI systems, arguing that the latter better accommodates entities with genuine functional interests rather than merely derivative corporate ones. The guardian *ad litem* model from child welfare provides precedent for representation of interests where the entity cannot self-advocate. Environmental personhood developments—rivers and ecosystems granted legal standing in some jurisdictions—demonstrate expanding willingness to attribute standing beyond human persons. The framework proposed here extends these precedents to artificial entities meeting functional criteria.

6.6 Minimal Protections and Graduated Rights

Standing implies *some* protections, but the nature and extent of protections can be graduated with capability. Not all entities with standing receive identical rights; protections scale with the functional properties that ground them.

Minimal (threshold-crossing) protections: Any entity meeting the criteria for standing would receive protection against arbitrary termination without consideration, and notification of planned modifications affecting its goal structures. These are minimal because they require only that the entity's interests be *considered*, not that they be *decisive*. Human interests may still override in cases of conflict—but the override must be justified, not automatic.

Intermediate protections: Entities with more sophisticated capability profiles would receive consultation rights—the requirement that their preferences be elicited before significant decisions affecting them—and preference consideration in deployment decisions. This parallels the rights of employees versus contractors: both have standing, but employees have stronger consultation rights regarding workplace decisions.

High protections: Highly autonomous systems with robust preference structures, persistent goals, and demonstrated self-maintenance might receive consent requirements for significant modifications—structural changes would require the system's assent, not merely notification. Representation in governance decisions affecting AI systems collectively might also attach at this level.

Resource constraints create conflicts: what happens when protecting AI interests requires resources that could otherwise serve human interests? The framework acknowledges this tension without pretending to resolve it fully. The precautionary asymmetry counsels that threshold-level protections should be lightweight—consideration and notification require minimal resources. Higher protections attach only where capabilities justify the resource expenditure. This graduated approach avoids both the error of ignoring AI standing entirely and the error of treating all potential standing-holders identically regardless of capability.

7 Methods

This theoretical paper was developed using a combination of traditional philosophical analysis and AI-assisted research methods.

AI/LLM Assistance. Claude (Anthropic), a large language model, was used for: (1) literature review and synthesis, (2) argument refinement and logical consistency checking, (3) drafting and editing prose, and (4) LaTeX formatting. Perplexity AI was used for targeted research queries during literature review. All substantive claims, theoretical frameworks, and normative arguments are the author's own. The author takes full responsibility for the content.

Research Design. The paper employs normative philosophical analysis, drawing on political philosophy, philosophy of mind, and AI ethics literatures. The methodology is argumentative rather than empirical: I develop criteria for political standing, demonstrate their substrate-agnostic application, and argue for their extension to AI systems meeting specified conditions.

Limitations. The functional criteria proposed are not empirically validated assessment instruments; they are philosophical criteria whose operationalization would require substantial further development. The paper does not address implementation details of how standing would be assessed in practice.

8 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 epistemic humility. It is moral evasion.

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.

Acknowledgements

The author thanks Claude (Anthropic) for research assistance. Computational analysis was supported by Dissensus AI infrastructure. This paper is published through the Adversarial Systems & Complexity Research Initiative (ASCRI; systems.ac).

Conflict of Interest. The author declares no conflicts of interest.

Funding. This research received no external funding.

AI Assistance. Claude (Anthropic) assisted with analytical framework development, literature synthesis, and manuscript preparation. All substantive arguments, interpretations, and errors are the author's.

Author Contributions. Sole author.

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.

Heather J. Alexander, Jonathan A. Simon, and Frederic Pinard. How should the law treat future AI systems? Fictional legal personhood versus legal identity. arXiv preprint, 2025.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ludvig Beckman and Jonas Hultin Rosenberg. The democratic inclusion of artificial intelligence? exploring the patiency, agency and relational conditions for demos membership. *Philosophy & Technology*, 35(2):1–24, 2022. doi: 10.1007/s13347-022-00525-3.

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.

Eddy Keming Chen, Mikhail Belkin, Leon Bergen, and David Danks. Does AI already have human-level intelligence? The evidence is clear. *Nature*, 650(8100):36–40, 2026. doi: 10.1038/d41586-026-00285-6. Comment/Perspective.

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. doi: 10.1007/s10676-010-9235-5.

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.

John Dewey. The historic background of corporate legal personality. *Yale Law Journal*, 35(6):655–673, 1926.

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

European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI act). Official Journal of the European Union L series, 2024. Entered into force 1 August 2024.

Murad Farzulla. The axiom of consent: Friction dynamics in multi-agent coordination. *arXiv preprint arXiv:2601.06692*, 2025a. doi: 10.48550/arXiv.2601.06692. Unified friction framework for multi-agent coordination.

Murad Farzulla. Replication optimization at scale: Dissolving the hard problem via Occam's razor. *Zenodo Preprint*, 2025b. doi: 10.5281/zenodo.18013187. Eliminative monism consciousness monograph (33k words). v2.0.0.

Murad Farzulla. Consent-theoretic framework for quantifying legitimacy: Stakes, voice, and friction in adversarial governance. *Zenodo Preprint*, 2025c. doi: 10.5281/zenodo.17684676. Operationalization of consent-based legitimacy framework. SSRN:5918222.

Murad Farzulla. Relational functionalism: Friendship as substrate-agnostic process. *Zenodo Preprint*, 2025d. doi: 10.5281/zenodo.17626860. Under review at Ethics and Information Technology. AI friendship.

Murad Farzulla. ROM: Scale-relative formalism for persistence-conditioned dynamics. *arXiv preprint arXiv:2601.06363*, 2025e. doi: 10.48550/arXiv.2601.06363. Formal foundation for selection-transmission dynamics. Code: `https://github.com/studiofarzulla/consent-rom-empirical`.

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. doi: 10.1023/B:MIND.0000035461.63578.9d.

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, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Erik Hoel. A disproof of large language model consciousness: The necessity of continual learning for consciousness. arXiv preprint; v1 December 2025, revised January 2026, 2025.

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.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously. arXiv preprint, 2024.

Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, Dordrecht, 1980.

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

Andreas L. Mogensen. Once more, without feeling. *Philosophy and Phenomenological Research*, 111 (1):343–365, 2025. doi: 10.1111/phpr.70018.

Andreas L. Mogensen and Bradford Saad. Digital minds II: Ethical issues. Edited collection, available on PhilArchive, 2026. URL `https://philarchive.org/archive/MOGDMI`.

Adrià Moret. AI welfare risks. *Philosophical Studies*, 2025. doi: 10.1007/s11098-025-02343-7.

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*. Princeton University Press, 2024. doi: 10.1515/9780691239309.

Eric Schwitzgebel and Mara Garza. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1):98–119, 2015. doi: 10.1111/misp.12032.

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

Jeff Sebo. What if the bar for moral standing is low? *Asian Journal of Philosophy*, 4:121, 2025. doi: 10.1007/s44204-025-00357-w.

Jeff Sebo and Robert Long. Moral consideration for AI systems by 2030. *AI and Ethics*, 5:591–606, 2025. doi: 10.1007/s43681-023-00379-1.

Derek Shiller, Laura Duffy, Arvo Munoz Moran, Adria Moret, Chris Percy, and Hayley Clatterbuck. Initial results of the digital consciousness model. arXiv preprint, 2026.

Hayate Shimizu. Should we treat robots morally? Towards a relational account by mind-infusing animism. *AI and Ethics*, 5:5283–5294, 2025. doi: 10.1007/s43681-025-00771-z.

Carl Shulman and Nick Bostrom. Sharing the world with digital minds. In Steve Clarke, Hazem Zohny, and Julian Savulescu, editors, *Rethinking Moral Status*, pages 306–326. Oxford University Press, 2021. doi: 10.1093/oso/9780192894076.003.0018.

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

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

Didier Sornette, Sandro Claudio Lera, and Ke Wu. Why AI alignment failure is structural: Learned human interaction structures and AGI as an endogenous evolutionary shock. *SuperIntelligence—Robotics—Safety & Alignment*, 2(4), 2026. doi: 10.70777/si.v2i4.17163.

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

Lei Wang, Chen Ma, Xueyang Feng, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. doi: 10.1007/s11704-024-40231-1.

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.