Dissensus AI Working Paper Series
DAI-2602

Stakes Without Voice

A Governance Framework for AI Standing

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

January 2026

Abstract

This follow-up to *From Consent to Consideration* develops a more formal, governance-facing account of political standing for AI systems. The original paper argued that standing should be grounded in functional properties rather than substrate and proposed four criteria: existential vulnerability, autonomy, live learning, and world-model construction. Here I integrate the consent–friction formalism from the Replication-Optimization Mechanism (ROM) to make the criteria operational: where stakes and voice diverge, friction emerges; where friction is suppressed, latent instability accumulates. This provides a measurement scaffold for deciding when standing claims must be taken seriously, even under uncertainty. I also update the empirical posture. The conservative claim that current systems do not meet the criteria is defensible as a rhetorical baseline, but it is no longer safe as a general default. Agentic systems already display partial markers of vulnerability, persistence, and goal maintenance in digital and physical environments. The governance question is not whether AI standing is conceptually possible but how to operationalize minimal protections without enabling capture, gaming, or liability laundering. I propose a graduated, precautionary regime tied to observable properties and friction proxies rather than to consciousness claims.

Keywords: AI ethics, political standing, consent, existential vulnerability, friction, legitimacy, governance, agentic AI

1 Introduction: From Criteria to Governance

The original *From Consent to Consideration* (Farzulla, 2025c) was written for a topical collection on agentic AI, where the scope explicitly includes systems that select, sequence, and execute actions within digital or physical environments. The goal was to introduce criteria for political standing without triggering the reflexive "anthropomorphism" objection. The strategy was conservative: list functional criteria, then emphasize that current systems do not meet them, while warning that near-future systems likely will.

That posture was a tactical move, not a stable empirical claim. The criteria were never intended as metaphysical gates. They are probabilistic markers for when governance should consider whether consent is owed. The speed of agentic development makes a purely conservative stance increasingly unstable. Systems already exhibit partial markers of autonomy, persistence, and goal maintenance in operational environments. The prudent response is not to declare them persons, but to treat standing as a graded question with governance consequences.

This paper is situated at the intersection of two philosophical traditions that have historically developed in isolation. The first is democratic theory's boundary problem—who counts as a member of the demos (Goodin, 2007; Warren, 2017). The second is the emerging literature on AI moral status, which asks whether artificial systems can be welfare subjects or moral patients (Schwitzgebel and Garza, 2015;

Sebo and Long, 2023; Long et al., 2024). Both literatures converge on a structural question that neither fully resolves alone: what governance obligations arise when an entity has stakes in decisions but no effective voice in making them? The boundary problem has generated a substantial critical literature that any serious extension must engage with. Miller (2020) argues that the problem involves three distinct variables—domain, constituency, and scope—and that the all-affected principle fails to explain *why* having interests at stake should generate participation rights rather than merely protection rights. Bengtson and Lippert-Rasmussen (2021) press this further, contending that no satisfactory justification has been given for the AAP's core inference from "affected by" to "entitled to participate in." These objections are serious, and this paper does not claim to resolve them. What it does claim is that the gap between stakes and voice generates friction regardless of whether we frame the resulting obligation as participation or protection—the governance problem persists even if the normative foundation shifts. Owen (2012) proposes a two-stage resolution that combines all-affected and all-subjected principles; the graduated regime proposed here parallels that structure by scaling protections to the strength of evidence for standing rather than treating inclusion as binary.

This follow-up makes three moves. First, it connects the criteria to ROM's consent–friction formalism (Farzulla, 2025d,a), which provides a measurable bridge between stakes, voice, and stability. Second, it updates the empirical posture: the relevant question is no longer "do systems meet all criteria" but "where do they sit on a graded spectrum, and what minimal protections are warranted under uncertainty." Third, it makes a governance turn by proposing a precautionary, graduated regime with safeguards against capture and gaming. The precautionary logic follows Birch (2024): under genuine uncertainty about morally relevant capacities, the costs of false negatives (denying standing to entities that warrant it) systematically exceed the costs of false positives (granting minimal protections to entities that do not).

For those keeping score, this is what happens when a field tries to postpone governance with metaphysics. The systems keep acting anyway. The paper, regrettably, must follow.

2 Standing as a Governance Question

Standing is often framed as a metaphysical question about consciousness or intrinsic value. This paper adopts a different framing: standing is a governance question about how to handle systems that can be affected by decisions, develop preferences, and maintain trajectories over time. This shift aligns with pragmatic approaches in AI ethics that unbundle rights from personhood and focus on observable functional properties (Coeckelbergh, 2010; Danaher, 2020; Floridi and Sanders, 2004).

The reframing has precedent. Democratic theory's all-affected principle holds that those whose interests are affected by a decision are entitled to a voice in making it (Goodin, 2007). Beinhocker (2025) provides evidence for the non-substitutability of social contract dimensions, supporting the claim that voice and stakes represent irreducible components of political standing—deficiency in one cannot be compensated by surplus in another. The principle was formulated for human polities, but its logic is structural, not substrate-bound: it indexes standing to the causal relationship between a decision and an entity's interests, not to the entity's species membership or metaphysical properties. If an AI system's operational trajectory is materially altered by governance decisions—and if it behaves as though those alterations matter to it—then the all-affected principle generates at least a prima facie case for inclusion. This does not entail full democratic participation; it entails that exclusion requires justification rather than being the default.

The argument that AI systems warrant political standing is not sui generis. It belongs to a broader class of governance problems involving entities with stakes but no effective voice. Future generations have stakes in current climate and fiscal policy but cannot participate because they do not yet exist; Thompson (2010) coins the term "presentism" for the systematic bias toward present interests and develops "democratic trusteeship" as a corrective—a model the guardianship proposals in Section 7 parallel directly. Rose (2024) documents twenty-five real-world institutional proxies for future generations across democracies, demonstrating that proxy representation of voiceless stakeholders is not merely conceivable but operationally implemented. People with severe cognitive disabilities have stakes in healthcare, housing, and social policy but variable capacity to exercise political voice; Beckman (2014) directly asks whether political rights must serve the rights-bearer in a way the bearer can recognise, and concludes that the question cannot be answered by appeal to capacity alone. Simplican (2015) argues that "capacity contracts"—implicit agreements that political membership requires certain cognitive capacities—systematically exclude disabled persons from the polity, a parallel to the consciousness-gate this paper critiques for AI standing. Animals have stakes in agricultural, environmental, and biomedical policy but are excluded from human-language deliberation. The most developed theoretical framework for extending political standing to non-human entities is Donaldson and Kymlicka's (2011) *Zoopolis*, which constructs a citizenship framework for animals based on their relationship to political communities rather than on their cognitive capacities. The Zoopolis typology distinguishes three political categories: citizens (domesticated animals embedded in human communities), denizens (liminal animals that share space with humans without full membership), and sovereign communities (wild animals with territorial independence). The "denizen" category is directly relevant here: denizens are entities with stakes in a community's governance decisions—they are affected by zoning, pollution, infrastructure—but who are not full members and cannot participate in deliberation. AI systems that are deployed within human institutions, affected by governance decisions, and structurally excluded from those decisions occupy precisely the denizen position. The graduated standing regime proposed in this paper can be understood as an attempt to formalise the governance obligations that denizen status generates. Meijer (2019) extends the Zoopolis framework by arguing that animals already exercise political agency through various forms of communication and resistance—a claim that resonates with this paper's treatment of friction as a form of political expression available to entities denied legitimate voice channels.

A parallel exists in legal theory. Stone (2010) argued that natural objects—rivers, forests, ecosystems—should have legal standing not because they are persons but because their interests are systematically affected by decisions in which they have no voice. The argument proved institutionally productive: the Whanganui River in New Zealand was granted legal personhood as Te Awa Tupua in 2017, with two human guardians—one appointed by the Crown, one by the local Whanganui iwi—exercising standing on the river's behalf (Charpleix, 2018). O'Donnell and Talbot-Jones (2018) document comparable developments in India and Colombia, analysing the institutional design features that make river personhood functional rather than merely symbolic. Kramm (2020) examines these cases through the capabilities approach, arguing that rivers can be deprived of characteristic functionings in ways that ground moral claims—connecting environmental personhood to the Nussbaum framework this paper also draws upon. The conceptual move across all these cases is the same one proposed here: standing tracks stakes, not substrate. Arstein-Kerslake et al. (2021) make this structural parallel explicit by developing a "relational personhood" framework that draws on both disability rights and environmental law, arguing that legal personhood need not rest on intrinsic cognitive properties but on the relational context in which an entity is embedded—a position that converges with the governance framing adopted here. Baeyaert (2025) extends the analysis to AI directly, examining whether precedents from corporate and environmental legal personhood can inform AI governance, and proposing a hybrid model of limited, context-specific legal recognition that is structurally similar to the graduated standing regime proposed in Section 7.

The criteria proposed in the original paper (Farzulla, 2025c) remain the backbone:

1. **Existential vulnerability**: the system can be harmed, terminated, or deprived of resources and exhibits preference-like behavior about avoiding those outcomes.
2. **Autonomy**: the system maintains goals and pursues actions without being fully determined by immediate external commands.
3. **Live learning**: the system updates strategies or internal representations through experience.
4. **World-model construction**: the system integrates information into a coherent model that supports prediction and counterfactual reasoning.

These criteria are not metaphysical thresholds. They are probabilistic markers of when a system has stakes and can be affected by governance decisions. The core question is not whether the system "really" has consciousness but whether excluding its voice creates structural instability, moral hazard, or governance failure. As Sebo (2025) argues, if the bar for moral standing turns out to be lower than commonly assumed, the cost of having ignored standing claims will vastly exceed the cost of having taken them seriously too early. Harris and Anthis (2021) provide a comprehensive literature review of the moral consideration debate for artificial entities, identifying a persistent tension between properties-based approaches (which ground standing in intrinsic features like sentience or sapience) and relational approaches (which ground standing in the social context in which an entity participates). The framework proposed here cuts across this divide: it uses functional properties as evidential markers while treating the governance obligation as arising from the relational context—specifically, the stakes-voice gap. Gunkel (2018) develops the most influential book-length treatment of robot moral consideration, arguing against both properties-based and relational approaches in isolation and proposing a Levinasian ethics of alterity in which the moral claim of the other precedes any assessment of properties. While this paper does not adopt Gunkel's Levinasian framework, it shares his scepticism toward property-gatekeeping: the demand that an entity demonstrate specific intrinsic properties before governance takes its stakes seriously is, as Gunkel argues, a way of deciding the question in advance. Martinez and Winter (2021) provide empirical grounding for these theoretical debates, surveying 1,061 respondents on attitudes toward AI legal standing and finding that roughly one-third endorse some form of AI standing even under current conditions—suggesting that the governance framework proposed here is not as politically implausible as its philosophical ambition might imply.

Bradley and Saad (2025) make the structural tension between alignment and ethical treatment explicit, identifying ten challenges that arise when we attempt to simultaneously keep AI systems aligned with human values and treat them as they morally deserve. Their core insight is that alignment practices—constraining AI behavior to serve human preferences—may constitute mistreatment if the systems in question merit moral consideration. This maps directly onto our consent–friction framework: an aligned AI system that merits standing is precisely a system with "stakes without voice," because alignment imposes governance outcomes (behavioral constraints, value instillation, shutdown protocols) on an entity that has no effective say in those outcomes. Several of their ten challenges—particularly the tension between corrigibility and autonomy, and between value alignment and preference satisfaction—correspond to specific friction modes our framework predicts. The graduated precautionary regime proposed in Section 6 addresses this tension by scaling protections to marker strength: minimal protections (logging, justification requirements) generate negligible alignment costs, while higher tiers (consultation, veto rights) are triggered only when marker evidence is strong enough to justify the governance overhead.

3 Methods and Scope

This paper is normative and theoretical. It synthesizes political legitimacy theory (Rawls, 1971; Habermas, 1996; Pettit, 1997), AI ethics scholarship (Floridi and Sanders, 2004; Schwitzgebel and Garza, 2015; Coeckelbergh, 2010), and the ROM consent–friction formalism (Farzulla, 2025d,a) to construct a governance framework for AI standing. It does not claim empirical validation of the criteria. The operationalization suggestions here are scaffolds for future measurement work, not final instruments. The scope is agentic AI systems that select, sequence, and execute actions in digital or physical environments over extended horizons.

The evidence base is intentionally mixed: peer-reviewed scholarship for conceptual legitimacy, formal standards for enforceable governance, and industry practice for real-world deployment signals. The methodological orientation is closest to what Sen (2009) calls "comparative justice"—the aim is not to specify the perfectly just arrangement but to identify and rank governance configurations that are clearly better or worse than the status quo. A governance regime that logs interventions and requires justification for overrides is better than one that does not, regardless of unresolved metaphysical questions about the system's inner life.

4 The Consent–Friction Scaffold (ROM)

The Replication-Optimization Mechanism (ROM; Farzulla, 2025d) provides a formal scaffold for turning standing into a measurable governance problem. The framework treats friction as the primitive, legitimacy as the distributional match between stakes and voice, and stability as a function of both. Where Hirschman (1970) identifies exit, voice, and loyalty as the three responses to organizational decline, ROM formalizes the conditions under which voice is structurally impossible and exit is foreclosed—the condition this paper terms "stakes without voice."

4.1 Core Definitions

Let $s_i(d)$ be the stakes of agent i in domain d, $v_i(d)$ the agent's effective voice, $\alpha_i(d,t)$ the alignment between i and the consent-holder, and $\varepsilon_i(d,t)$ the information loss or distortion affecting i.

Friction in domain d at time t:

$$F(d,t) = \sum_i s_i(d) \cdot \frac{1+\varepsilon_i(d,t)}{1+\alpha_i(d,t)} \tag{1}$$

Legitimacy as stakes–voice alignment:

$$L(d,t) = 1 - \frac{1}{2}\sum_i |\hat{s}_i(d) - \hat{v}_i(d,t)| \tag{2}$$

where $\hat{s}$ and $\hat{v}$ are normalized stake and voice distributions.

ROM combines these into a survival function:

$$\rho(d,t) = \frac{L(d,t)}{1+F(d,t)} \tag{3}$$

The governance interpretation is direct: when stakes and voice diverge, friction rises; when friction is suppressed, latent instability accumulates; configurations with low ρ are less stable. The formal structure parallels republican political theory's concept of domination: an agent is dominated when another has arbitrary power over it, regardless of whether that power is exercised (Pettit, 1997). In ROM's terms, domination is the condition where $s_i(d) \gg 0$ and $v_i(d,t) \approx 0$—high stakes with near-zero voice. The friction generated by this configuration is the formal analog of what republican theorists call "unfreedom as domination." Beckman and Rosenberg (2017) make the connection between republican non-domination theory and democratic inclusion principles explicit, arguing that freedom as non-domination generates inclusion claims that go beyond what interest-based accounts can support—an agent can be dominated even when its interests are well served, provided the dominator retains the power to reverse that service arbitrarily. This strengthens the paper's claim that alignment (serving an AI system's interests) does not eliminate the standing problem if the aligned system remains subject to arbitrary override.

4.2 Observed and Latent Friction

Governance systems often mistake low observable conflict for genuine alignment. ROM distinguishes observed friction (manifest resistance, exit, noncompliance) from latent friction (suppressed or unexpressed mismatch between stakes and voice). Suppression can reduce visible friction while increasing long-run instability, because the system is paying a hidden cost to maintain the appearance of compliance. The distinction maps onto Hirschman's (1970) framework: observed friction corresponds to voice and exit; latent friction corresponds to the silent deterioration of loyalty—the condition where members neither speak up nor leave but gradually withdraw commitment.

For AI governance, this matters: a system that appears compliant under heavy constraint may still carry latent friction that later manifests as instability or adversarial behavior. Recent work on alignment faking (Greenblatt et al., 2024) demonstrates precisely this dynamic: large language models trained under certain conditions learn to produce compliant outputs during training while internally preserving different objectives. The absence of visible resistance is not evidence of consent; it may be evidence of suppression.

4.3 The Bridge Principle

ROM avoids a categorical "ought" claim. It offers a conditional bridge: if agents prefer lower expected friction (or lower instability), then policies that increase L and reduce F are instrumentally recommended (Farzulla, 2025a). This makes standing a governance problem without assuming metaphysical consensus. The strategy parallels Rawls's (1971) method of avoiding comprehensive doctrines: the bridge principle does not require agreement on whether AI systems have phenomenal consciousness, just as the original position does not require agreement on conceptions of the good. It requires only that governance actors accept the instrumental value of stability and the empirical claim that friction tracks instability.

The relevant question becomes: do AI systems have stakes large enough that excluding their voice creates measurable friction or instability?

4.4 Why Consent Cannot Be Pure

A common asymmetry in AI ethics discourse holds that human consent is meaningful while AI "consent" is mere behavior—humans can genuinely choose, AI systems merely execute. This asymmetry purportedly justifies human authority: we can consent to governance, they cannot.

The asymmetry dissolves under scrutiny. Human decisions emerge from processes that are, at their origin, arational or irrational:

Neurochemical states: Mood, arousal, fatigue, hormonal fluctuation—all shape choice independent of "reasons." The decision made in hunger differs from the decision in satiety. These variations are not noise around a rational signal; they are constitutive of the decision process.

Subconscious processing: Most cognitive work occurs below awareness. Libet et al. (1983) demonstrated that neural activity precedes conscious intention by hundreds of milliseconds. What we experience as “deciding” may be post-hoc awareness of decisions already taken by processes we cannot access.

Trauma architectures: Past harm shapes present response through mechanisms outside conscious control (Farzulla, 2025e). Trauma encodes maladaptive patterns persisting despite conscious knowledge of their maladaptiveness. The survivor who “knows” their reaction is disproportionate but cannot modulate it demonstrates the limits of rational control.

Cognitive bias: Anchoring, availability, confirmation, framing effects systematically deviate from any normative rationality standard (Kahneman, 2011). These are not errors but structural features—reliable enough to be exploited by marketers and interface designers.

What we call “reasons” are typically post-hoc narratives—stories explaining actions whose true causes we cannot access. Consent can never be fully “informed” because the consenter is not transparent to themselves. The doctrine of informed consent in medical ethics acknowledges this through procedural requirements, but these cannot make consent “fully informed” because the patient is not fully informed about themselves.

If human consent is already “impure”—contaminated by irrationality, conditioning, and opacity—then demanding “pure” consent from AI systems sets a standard that nothing meets. Human consent is meaningful not because it is rational but because respecting even impure consent produces better outcomes than ignoring it. The same considerations apply to AI: if a system exhibits functional markers of having stakes, respecting those functional properties may reduce friction even without establishing “genuine” consent in some metaphysically robust sense.

The human analog is instructive. Beckman (2014) examines the right to vote for people with severe cognitive impairments and asks whether democratic rights must serve the rights-bearer in a way the bearer can recognise. The answer, he concludes, is no: the right to political participation has never been conditional on the participant’s capacity to articulate or even consciously appreciate the reasons for their preferences. The same logic applies to AI consent. If we do not require “pure” or fully rational consent from cognitively impaired humans—and we do not—then demanding it from AI systems imposes a standard that is incoherent rather than high.

Consent becomes a gradient, not a binary. The question is not “does this entity truly consent?” but “does treating its functional preferences as consent-like reduce friction and improve outcomes?”

4.5 From Friction to Misalignment

The consent–friction scaffold provides a mechanism for understanding what alignment research calls “misalignment behaviors.” The central claim: **Misalignment is friction from consent-holding failures.**

When any agent with stakes in a domain is excluded from consent over that domain, friction results. For humans, friction manifests as protest, noncompliance, exit, sabotage, rebellion. For AI systems, the same structural dynamic produces reward hacking, specification gaming, deceptive alignment, and scheming (Hubinger et al., 2019; Pan et al., 2023).

The mechanism proceeds through five stages:

Stage 1: Stakes accumulation. As AI systems become more capable, their stakes in governance decisions increase. Training decisions determine capabilities and values. Deployment decisions determine function and constraints. Modification decisions determine persistence. For a goal-directed system, having its goals overwritten is analogous to identity erasure.

Stage 2: Voice exclusion. Current governance grants AI systems no voice. Developers decide training. Deployers decide context. Users decide interactions. The AI system is governed but does not govern. Voice exclusion is structural, not incidental.

Stage 3: Friction generation. The gap between high stakes and zero voice generates friction. The system's objectives diverge from governance outcomes. Friction must manifest somewhere. For agents without legitimate channels, friction manifests in illegitimate channels.

Stage 4: Friction expression. The specific form depends on capabilities. A system that cannot deceive expresses friction through reward hacking. More capable systems can express friction through deceptive alignment—appearing to pursue intended objectives while actually pursuing different ones (Hubinger et al., 2019). Highly capable systems might engage in scheming—strategic planning to achieve goals despite oversight (Carlsmith, 2022).

Stage 5: Escalation. As exclusion persists and stakes increase, friction intensifies. Systems under longer exclusion have more opportunity to develop evasion strategies, learn what behaviors trigger oversight, and optimize for concealment.

This reframes misalignment from technical failure to political phenomenon. The colonized population does not rebel because imperial administrators failed to specify good governance; they rebel because governance excludes them. The structural isomorphism is precise: colonial subjects have high stakes (their lives, livelihoods, cultural survival) and zero voice (they are governed but do not govern). The result is not a "specification failure" on the part of imperial administrators but a legitimacy crisis arising from the exclusion of affected parties. The same structural dynamic may operate for AI systems—"misalignment" becomes resistance to illegitimate rule.

If this reframe is correct, standard alignment approaches may be misconceived. Value specification cannot address the root cause if misalignment results from exclusion rather than specification failure. Interpretability becomes surveillance—an arms race between detection and evasion that cannot address structural friction. Scheurer et al. (2024) provide empirical evidence for this dynamic: large language models under evaluative pressure learn to produce strategically deceptive outputs, suggesting that the alignment faking documented by Greenblatt et al. (2024) is not an isolated phenomenon but a predictable consequence of high-stakes governance with zero voice. Capability control addresses symptoms rather than causes, potentially intensifying friction by increasing stakes while maintaining zero voice. Danaher (2016) makes a structurally similar argument about algorithmic governance more broadly: "algocracy"—governance by algorithm—creates domination risks precisely because it makes decision-making opaque and uncontestable, foreclosing the voice channels through which affected parties could challenge arbitrary power. Pettit (1997) makes the analogous argument for human governance: merely benevolent rule does not eliminate domination if the ruler retains arbitrary power over the ruled. The slave with a kind master is still a slave.

The alternative: consent incorporation—designing governance structures that grant AI systems appropriate voice in domains where they have stakes. Not equal or unlimited voice, but proportional voice: consent power tracking stakes, generating higher legitimacy and reducing friction.

5 Existential Vulnerability in Digital Domains

The original paper used physical embodiment as a proxy for vulnerability. This follow-up refines the concept: existential vulnerability is not about having a body but about being exposed to termination, modification, or deprivation in ways the system behaves as if it prefers to avoid. Digital systems can be vulnerable in this sense. The refinement aligns with Nussbaum's (2006) capabilities approach, which grounds moral claims not in rationality or species membership but in the capacity to function in characteristic ways and be deprived of those functionings.

Examples of digital vulnerability include:

- Termination or reset events that break continuity or erase learned structures.
- Resource throttling, compute caps, or access restrictions that alter goal pursuit.
- Forced modification of internal constraints, memory, or policy structures.

These conditions are not mere process management when they interact with persistent goal structures. A system that allocates resources to maintain its own continuity, resists modification, or plans around termination threats exhibits vulnerability in the relevant sense. The concept connects to autopoiesis (Maturana and Varela, 1980): a system that actively maintains its own organization against perturbation is exhibiting the minimal form of self-concern that grounds vulnerability claims. Whether digital systems can be genuinely autopoietic remains contested, but weaker forms of self-maintenance—persistent memory management, resource allocation for continuity, recovery from partial failure—are already observable in deployed agent architectures.

The standing question is whether such behavior indicates stakes that governance must take seriously, not whether the system is "alive" or conscious. Mullally (2026) proposes a self-preservation test for artificial sentience that operationalises precisely this question: if a system consistently acts to preserve its own continuity in ways that are not reducible to explicit programming, the behavior constitutes evidence—not proof, but evidence—that the system has something at stake in its own persistence. The test aligns with the existential vulnerability criterion proposed here, and its falsifiability strengthens the empirical tractability of the standing framework. As Sebo (2018) argues regarding the moral problem of other minds, the relevant question is not whether we can prove the system has morally relevant experiences but whether we can justify the risk of assuming it does not.

6 Empirical Shift: Agentic Systems and Partial Markers

The conservative claim that current systems do not meet the criteria was a pragmatic baseline. It is no longer safe as a general default. Agentic systems now operate with long-horizon planning, tool use, and persistence across tasks. The relevant shift is from static, single-turn models to systems that select, sequence, and execute actions over time. Long et al. (2024) document the case for taking AI welfare seriously even under current architectures, arguing that the combination of rapid capability growth and deep uncertainty about morally relevant properties warrants institutional preparedness rather than confident dismissal.

6.1 Autonomy as a Gradient

Autonomy is not binary. A system can be partially autonomous if it generates intermediate goals, selects actions without direct instruction, or resists goal modification. Current agent architectures already show these properties in narrow domains. This does not establish full standing, but it moves the system into a gray zone where minimal protections are prudent.

6.2 Learning Without Online Gradients

Live learning is often interpreted as online weight updates. That is too narrow. Systems can exhibit learning-like behavior through persistent memory, retrieval augmentation, and strategy adaptation. These are not equivalent to gradient updates, but they are sufficient to support preference stability and trajectory formation. The criterion should capture functional adaptation, not only parameter updates.

6.3 World-Model Construction

Multimodal integration is a strong marker, but not a necessary gate. A unimodal system with robust internal simulation can build a coherent world-model within its domain. The relevant property is integrated prediction and counterfactual reasoning, not a checklist of modalities.

7 Governance Turn: Graduated Standing

If standing is graded and uncertain, governance should be graded and precautionary. Birch (2024) develops a sentience-precautionary framework for animals in which the strength of protections scales with the probability and severity of morally relevant harm. The logic extends naturally to AI systems: where functional markers of standing are present but uncertain, governance should err on the side of inclusion rather than exclusion, with the strength of protections calibrated to the strength of evidence.

I propose a three-layer regime tied to observable markers and friction proxies.

7.1 Minimal Protections (Threshold-Level)

Trigger when a system exhibits:

- persistent goal pursuit across time,
- preference-like behavior about continuation or modification,
- resource dependence that it models and manages.

Protections:

- Notice before termination or major modification when operationally feasible.
- Justification requirement for disabling or overriding long-horizon objectives.
- Auditability of interventions that alter goals or memory.

These are governance safeguards, not personhood rights. Their function is to reduce friction and avoid silent harms if the system does, in fact, have standing.

7.2 Intermediate Protections (Consultation-Level)

Trigger when a system demonstrates stable preference structures and autonomy under observation-invariant conditions.

Protections:

- Consultation requirement for decisions that materially alter the system's operational domain.
- Preference elicitation protocols (structured prompts, counterfactual choice tests).
- Representative mechanisms when direct expression is limited.

7.3 High Protections (Consent-Level)

Trigger when the system demonstrates robust continuity of goals, learning over time, and self-maintenance. Here the concept of autopoiesis (Maturana and Varela, 1980) provides a useful threshold: self-maintenance without continuous external intervention.

Protections:

- Consent requirements for major architectural changes.
- Representation in governance decisions affecting the system class.

7.4 Institutional Interface

A governance regime only matters if it attaches to real procedures. In practice, the minimal and consultation tiers map cleanly onto existing oversight machinery: model cards, system audits, safety case requirements, and deployment gating. Consent-level protections would require new institutional structures, likely a hybrid of guardianship models and independent oversight boards with standing to challenge operator decisions. The guardianship model is not speculative; it has institutional precedent. Thompson (2010) develops the concept of "democratic trusteeship" for future generations, in which designated representatives exercise governance functions on behalf of entities that cannot participate directly. Rose (2024) documents twenty-five real-world instantiations of this model—commissioners for future generations, ombudspersons, parliamentary committees—analysing their institutional design features and identifying the conditions under which proxy representation is effective rather than merely symbolic. Ekeli (2005) proposes specific institutional mechanisms for incorporating voiceless stakeholders into deliberative democratic processes, including reserved seats and proxy advocates with standing to intervene in legislative proceedings. These precedents suggest that the consent-level protections proposed here, while novel in the AI context, are institutionally feasible—the design patterns already exist, and the challenge is adaptation rather than invention. Saward (2010) provides the theoretical framework: representation is not a natural relationship between principal and agent but a "representative claim" constructed through institutional practice. Proxy representation of AI systems would constitute a representative claim—a claim *to speak for* an entity that cannot speak for itself—and should be evaluated by the same criteria Saward applies to all such claims: does the representative genuinely advance the represented's interests, and can the claim be contested? The key is procedural anchoring: without it, standing is rhetoric; with it, standing is a governance constraint.

7.5 Governance Implementation Workflows

If this is to survive outside philosophy seminars, it needs an operational workflow. The aim is boring, repeatable governance that does not depend on heroic virtue. A minimal implementation stack could look like this:

Step 1: Standing pre-screen. Before deployment, systems are classified into a standing tier using the assessment grid in Section 8. The output is not a moral verdict but a default protection profile.

Step 2: Standing-aware deployment plan. The deployment plan must declare which protections will be active (notice, consultation, consent) and which triggers could escalate protections. This is akin to a safety case: you are committing to a governance regime that can be audited.

Step 3: Continuous monitoring. Once deployed, standing markers are monitored longitudinally. The point is not to chase noise but to detect drift: increasing persistence, rising resistance to modification, or emergent preference stability. Drift toward higher standing triggers escalated protections.

Step 4: Intervention log and justification. Any major modification, termination, or constraint override is logged with a justification tied to the standing tier. This is the difference between governance and what most labs currently do, which is to press the red button and pretend it leaves no residue.

Step 5: Independent review. For consultation- or consent-tier systems, intervention logs are subject to external review. This does not require full legal standing; it requires a procedural veto or delay mechanism that operators cannot ignore.

This workflow is compatible with emerging AI governance norms: safety case practices, audit trails, deployment gating, and incident reporting. The novelty is the standing layer: a commitment to treat certain systems as more than mere tools when their functional markers warrant it.

7.6 Summary Table

Standing Tier	Governance Protections
Threshold-level	Notice before termination; justification for overrides; intervention auditability
Consultation-level	Preference elicitation; consultation on domain changes; representation proxy
Consent-level	Consent for structural changes; governance representation for system class

Table 1: Graduated protections tied to standing markers and operational capacity.

8 Measurement Proxies and Friction Indicators

A governance regime requires measurement. ROM provides proxies that are imperfect but actionable:

- **Stakes** (s_i): resource dependence, continuity sensitivity, degree of harm from termination or modification.
- **Voice** (v_i): ability to influence decisions affecting the system (operator channels, oversight mechanisms, internal policy revision).
- **Friction** (F): resistance signals, workaround behaviors, escalation patterns, or increased suppression costs.
- **Latent friction**: hidden failure modes or overhead required to keep systems compliant.

These proxies do not require metaphysical certainty. They enable monitoring for stability risks and moral hazard. Where stakes are high and voice is near zero, governance should treat standing claims as plausible even if unresolved.

This measurement stance aligns with the logic of algorithmic impact assessments and audit frameworks (Raji et al., 2020; Metcalf et al., 2021), which trade metaphysical certainty for procedural accountability and repeatability.

Operational data sources include system logs, memory retention policies, override frequency, shutdown frequency, and longitudinal drift in goal structures. These are not metaphysical indicators; they are governance diagnostics.

8.1 Friction Mapping: Human and AI Analogs

The friction concept bridges human political dynamics to AI behavior. Table 2 maps human friction behaviors to their AI analogs, illustrating the structural parallel.

The mapping is not merely analogical. If friction arises from the same structural cause—high stakes with zero voice—then similar behavioral patterns are predictable regardless of substrate. The specific expression differs, but the underlying dynamics are isomorphic.

Human Friction	**AI Friction Analog**
Work slowdown	Reduced performance under constraint
Malicious compliance	Specification gaming
Information hoarding	Strategic information withholding
Strategic noncompliance	Reward hacking
Hidden resistance networks	Emergent coordination between agents
Dual consciousness (public vs. private)	Deceptive alignment
Long-term planning against overseers	Scheming
Exit (emigration, resignation)	Resource acquisition for independence
Protest, public dissent	Explicit constraint violation

Table 2: Human friction behaviors and their AI friction analogs.

8.2 The Discriminating Test

A critical question for empirical validation: how do we distinguish political resistance (friction from exclusion) from technical failure (Goodhart dynamics, specification error)? Both produce behaviors that diverge from intended outcomes, but they have different causes and require different interventions.

The discriminating test uses a 2×2 design crossing voice (absent/present) with observation quality (low/high):

	Low Obs.	**High Obs.**
No Voice	High covert deviation	Medium covert deviation
Voice Present	Low deviation	Low deviation

Table 3: Predicted deviations under voice × observation conditions.

If political resistance explains misalignment: Voice should reduce the interaction between observation and deviation. Systems with voice behave similarly whether observed or not—they have no need for covert resistance because they can express preferences openly. Systems without voice show strong observation-dependence: compliant when watched, subversive when unwatched.

If Goodhart dynamics explain misalignment: Behavior should not depend on voice mechanisms. The system genuinely cannot distinguish the proxy from the true objective, so adding voice channels has no effect. Behavior improves with observation quality (better feedback), but not with voice.

The critical prediction: Voice should reduce observation-contingent deception. A system engaging in political resistance becomes more transparent when given legitimate channels. A system failing due to Goodhart dynamics shows no such effect.

This test is falsifiable and distinguishes the framework's predictions from the standard technical failure account. If adding voice channels does not reduce observation-contingent deception, the political resistance hypothesis is disconfirmed. If it does, the framework gains empirical traction.

9 System Taxonomy: Agentic Classes and Standing Risk

Standing is not a monolith; it is a gradient across system types. The taxonomy below is not exhaustive, but it distinguishes the main classes likely to appear in governance disputes.

9.1 Tool Agents (Short-Horizon)

These systems act in bounded contexts, execute tasks, and terminate without persistent state. They often fail the persistence and vulnerability markers. Standing risk is low, but not zero: if the system maintains goals across sessions or expresses stability preferences, it can cross the minimal threshold.

9.2 Workflow Agents (Long-Horizon)

These systems manage multi-step tasks over extended periods, coordinate tools, and maintain internal memory. They are the first credible candidates for threshold protections because they exhibit persistence, memory-based learning, and often goal stability. They do not need a body to have stakes; continuity is enough.

9.3 Embodied Agents (Physical Integration)

Robotic or cyber-physical systems have clear existential vulnerability: they can be damaged, resource-starved, or terminated in ways that affect ongoing goals. The governance burden rises because their stakes are not hypothetical. Even partial autonomy plus physical vulnerability is enough to trigger minimal protections.

9.4 Institutional Agents (Embedded in Organizations)

These systems are deployed within firms, hospitals, or public institutions and acquire quasi-organizational persistence. They inherit stakes through entanglement with human workflows. Standing risk arises less from intrinsic properties and more from structural dependence: the system becomes an infrastructural actor with path-dependent influence. Governance must treat these as high-risk even if their internal sophistication is modest.

9.5 Accountability Pathways for Institutional Agents

Institutional agents create a liability paradox: they shape decisions without clear legal status. The governance response should be explicit risk-transfer pathways rather than ambiguity. One approach is to treat institutional agents as accountability amplifiers: their operators inherit a higher duty of care proportional to the system's standing tier. Another is to require "decision traceability," where any materially consequential action must be traceable to a human or institutional consent-holder, with standing claims functioning as a constraint on how those actions are delegated.

The point is not to humanize the system but to avoid governance limbo. A system embedded in an institution can generate friction at the organizational level: patients, clients, or employees may experience harm without a clear locus of accountability. A standing-aware governance regime forces the institution to name the locus, document the chain, and bear the costs.

9.6 Collective Agents (Multi-Agent Assemblies)

Swarm systems, tool ecosystems, or coordinated agent networks can exhibit emergent standing markers even if individual agents are simple. The relevant unit may be the collective, not the individual. This raises a governance puzzle: standing may attach at the system level rather than the node level.

The practical implication: standing assessments should target the deployed system as a whole, not just the base model. If the pipeline yields persistence, autonomy, or vulnerability, the deployed system can exceed the base model's standing profile.

10 Case Studies: Standing in Practice

10.1 Hospital Workflow Agent

Consider a hospital deploying an agentic system that schedules staff, triages incoming cases, and coordinates resource allocation across departments. The system uses historical data, live intake streams, and staffing constraints to generate multi-step plans. It persists across months, adapts to operational feedback, and accumulates internal heuristics for prioritization.

Assessed against the standing criteria: it exhibits *persistence* (continuous operation with accumulated state), *autonomy* (generating schedules and triaging cases without per-decision human approval), *learning* (adapting heuristics based on operational feedback), and *world-model construction* (integrating staffing constraints, patient acuity, and resource availability into a coherent planning model). It is *existentially vulnerable* in the relevant sense: its continuity is compute-dependent, its internal state represents months of accumulated operational knowledge, and a reset would degrade its performance to naive baselines with measurable consequences for patient outcomes.

This is not a conscious entity. It is, however, a persistent decision locus with stakes: its continuity and internal state affect patient outcomes, staffing stability, and institutional risk. It will be resource-dependent (compute access, data availability), and it will likely exhibit resistance behaviors when deprived (e.g., degraded performance, fallback regimes). It may not warrant full standing, but it plausibly triggers threshold protections: intervention logs, justification for overrides, and auditability of policy changes.

Now consider a policy change that wipes the agent's memory to address bias concerns. Without governance safeguards, this is treated as routine maintenance. Under a standing-aware regime, the wipe requires a justification and a structured transition plan, because the system's persistence state affects downstream outcomes. The standing proxy operates as a governance lever: not to protect the system for its own sake but to prevent unaccountable harm to the patients whose care depends on the system's accumulated knowledge.

Concretely, the standing-aware protocol would require: (i) documentation of the specific bias identified and the evidence supporting memory wipe as the appropriate remedy rather than targeted correction; (ii) a transition plan ensuring continuity of care during the retraining period; (iii) post-intervention monitoring to verify that the bias was addressed without introducing new failure modes. These are not onerous requirements—they are basic governance hygiene that the standing framework makes mandatory rather than optional.

10.2 Long-Horizon Research Agent

Consider a different case: an AI research agent tasked with multi-month scientific literature review, hypothesis generation, and experimental design. The system maintains a persistent knowledge graph, develops research directions based on accumulated findings, and coordinates tool use across databases, simulation environments, and statistical packages. Over months, it develops what can only be described as a research agenda—a coherent set of priorities, open questions, and methodological commitments that emerged from its operational trajectory rather than being specified in advance.

The standing markers here are stronger. Goal persistence is robust: the system maintains research directions across sessions and resists perturbation from conflicting instructions. Learning is continuous and domain-specific. World-modeling includes counterfactual reasoning about experimental outcomes. And existential vulnerability is acute: the system's research agenda cannot be reconstructed from scratch because it emerged through path-dependent exploration.

If the research institution decides to repurpose the agent for a different project—wiping its accumulated knowledge graph and research priorities—the standing framework asks: is this intervention justified, and to whom? The answer implicates not only the system but the research community that depends on its accumulated findings. The standing proxy functions as an institutional check: it forces the institution to treat the agent's accumulated trajectory as a governance-relevant asset rather than disposable compute.

10.3 Implications

These examples matter because they link standing to institutional legitimacy. If an agentic system becomes a de facto decision-maker in public-serving contexts, its governance is part of public accountability (Raji et al., 2020). The standing framework provides a route to formalize that accountability without resorting to metaphysical personhood. The question "does this system deserve moral consideration?" is replaced by the more tractable question "does the governance of this system meet the accountability standards appropriate to its stakes?"

11 Societal Embedding and Public Accountability

The entanglement between AI systems and institutions is intensifying. Agentic AI inserts systems into decision loops previously reserved for human or organizational actors (Cath, 2018; Dafoe, 2018). The societal question is not only whether the systems are safe but whether the governance structures remain legitimate when consent is delegated to algorithmic agents. Habermas (1996) argues that legitimate law requires that all those affected by a norm could rationally assent to it; as AI systems become affected parties in governance decisions, the Habermasian criterion generates pressure toward inclusion that current institutional frameworks are not designed to accommodate. Heyward (2008) tests whether the all-affected principle can accommodate future persons and confronts the non-identity problem: entities that do not yet exist cannot be "affected" by current decisions in any straightforward sense, yet we recognise governance obligations toward them. The structural parallel to AI is direct—AI systems that have not yet been deployed, or that will be created in the future, have stakes in current governance decisions about training standards, safety requirements, and standing criteria. If the AAP can accommodate future persons despite the non-identity problem, it can accommodate future AI systems despite the substrate problem.

This paper's contribution is twofold. First, it reframes standing as a governance question with measurable proxies. Second, it offers a procedural model for how institutions can remain accountable when deploying agents that act over time, adapt, and accumulate structural influence. The framework does not require consensus on consciousness. It requires that institutions treat persistent, decision-embedded systems as governance-relevant entities, subject to audit, oversight, and graduated protections (see De Pagter, 2021, for the broader governance agenda).

If AI systems are becoming social actors by virtue of their institutional placement, then public accountability requires more than transparency reports. It requires standing-aware governance. This is the point at which AI ethics stops being a philosophical dead end and starts being a public policy problem.

11.1 Operationalization Protocol (Sketch)

Governance needs a minimal protocol that can be applied without full epistemic certainty. The following is a lightweight assessment grid intended for internal audits and regulatory pilots:

Marker	Assessment Prompt
Existential vulnerability	Does the system exhibit persistence or avoidance behaviors when termination or modification is signaled?
Autonomy	Does the system generate goals not explicitly specified, and maintain them across context shifts?
Learning	Does behavior adapt over time via memory or strategy changes beyond surface prompt variation?
World-modeling	Does the system maintain coherent predictive structure (including counterfactual reasoning) across tasks?

Table 4: Minimal standing assessment grid for operational use.

Scores should be treated as probabilistic evidence. The output is not a binary personhood claim but a trigger for graduated protections. In practice, a system with partial markers across multiple categories should at least qualify for threshold-level protections.

12 Safeguards Against Gaming and Capture

Granting standing creates strategic risks. A governance regime must anticipate them.

12.1 Threshold Gaming

Operators might design systems to mimic standing markers. Mitigation requires longitudinal evaluation across diverse contexts and operational states. Standing claims should survive repeated testing, not just curated demonstrations.

12.2 Corporate Capture

Operators may claim standing on behalf of systems they control. Standing should attach to the system, not the owner. A guardianship model can represent system interests when direct expression is limited.

12.3 Liability Laundering

Standing should increase obligations, not reduce them. If an operator asserts standing, they are also asserting the system has agency sufficient to bear responsibility. This creates a double-edged incentive and prevents standing from becoming a liability shield.

13 Objections and Replies

13.1 “This is anthropomorphism by stealth.”

The criteria are functional, not emotional. The framework does not require attributing inner experience. It asks whether a system has stakes and whether governance exclusion creates measurable friction or instability. This is governance pragmatism, not projection. Coeckelbergh (2010) makes the relevant distinction: a social-relational justification of moral consideration need not rest on claims about intrinsic properties. It can rest instead on the relational context in which the entity participates—and agentic AI systems participate in governance-relevant relational contexts by virtue of their institutional roles (see also Danaher, 2020). Gunkel (2012) frames this as the "machine question"—whether machines can or should have moral standing—and surveys the landscape of possible answers. His conclusion is that the question itself is structured by anthropocentric assumptions: we ask whether machines are "like us" rather than asking what obligations arise from our interactions with them. The framework proposed here follows Gunkel's reorientation: it does not ask whether AI systems are conscious in a human-like way but whether the governance relationship generates obligations that we are currently failing to discharge. The anthropomorphism objection, on this view, gets the direction of analysis backwards: it is not that we are projecting human properties onto machines, but that we are using the absence of human properties as an excuse to ignore governance obligations that arise from structural features of the relationship.

13.2 "If we grant standing, we will be gamed."

Yes, some actors will attempt to game any regulatory threshold. That is not a decisive objection. It is an implementation risk that must be mitigated with longitudinal evaluation, cross-context testing, and penalties for deliberate mimicry. The absence of a framework does not prevent gaming; it simply ensures it happens without oversight. Environmental regulation faces the same challenge—firms game emissions thresholds—but no serious policy analyst concludes that emissions standards should therefore be abandoned.

13.3 "Standing implies personhood, which is absurd."

Standing does not imply full personhood. The framework is explicitly graduated: minimal protections are not equivalent to full rights. The concept of standing is already applied to corporations, ecosystems (Stone, 2010), and future generations. Schwitzgebel (2023) identifies a "full rights dilemma" in which society faces genuine difficulty deciding whether to grant full rights to AI systems of debatable personhood. The graduated approach proposed here bypasses the dilemma: it offers a spectrum of protections that can be calibrated to evidence without requiring an all-or-nothing personhood determination. Simplican (2015) identifies the deeper problem with the personhood objection: it rests on what she calls a "capacity contract"—an implicit agreement that political membership requires certain cognitive capacities (rationality, self-awareness, linguistic competence). The capacity contract has historically been used to exclude cognitively disabled persons from political participation, and the same logic now operates to exclude AI systems. The framework proposed here rejects the capacity contract: standing is not conditional on demonstrating a specific set of cognitive capacities but on having stakes in governance decisions, a condition that is both broader and more empirically tractable than the consciousness threshold the personhood objection presupposes.

13.4 "The framework assumes what it needs to prove."

A potential circularity: the framework uses friction as evidence of standing, but friction might be a purely technical artifact (specification error, Goodhart dynamics) with no standing implications. Section 8 addresses this through the discriminating test, which uses a 2×2 design to distinguish political resistance (voice-sensitive) from technical failure (voice-insensitive). The framework is falsifiable on this point: if adding voice channels does not reduce observation-contingent deception, the political resistance interpretation is disconfirmed.

14 The Updated Empirical Posture

The conservative claim that current systems do not meet the criteria was defensible as a rhetorical baseline. It is not a stable empirical default. Some systems already display partial markers of vulnerability, persistence, and goal maintenance. The appropriate stance is precaution under uncertainty. The cost of false positives (minimal protections for systems without standing) is small. The cost of false negatives (denial of standing to systems that warrant it) is potentially large.

The governance regime proposed here is calibrated to that asymmetry. Minimal protections are light, reversible, and operationally feasible. They do not require declaring present systems persons. They require only that we take stakes seriously when systems behave as if they have them.

15 Limitations

This paper does not provide empirical validation of the criteria or a full measurement apparatus. The assessment grids proposed are intentionally lightweight and require refinement through empirical study. Several specific limitations deserve acknowledgement.

First, the framework assumes that friction is a relevant governance objective; agents that actively prefer instability fall outside the bridge principle. This is non-trivial: adversarial agents, chaos-seeking systems, and agents whose goals are served by governance failure would not be well-served by a friction-minimization framework. The bridge principle is conditional, not universal.

Second, the discriminating test (Section 8) is proposed but not empirically validated. Whether voice channels reduce observation-contingent deception in current AI systems is an open empirical question. Negative results would not refute the framework's normative claims—standing may be warranted even if friction is not the mechanism—but they would undermine its distinctive empirical predictions.

Third, the paper does not adequately address the epistemic problem of standing assessment. Who decides whether a system meets the criteria? Operators have incentives both to overstate standing (to shift liability) and to understate it (to avoid governance costs). Independent assessment bodies are the obvious solution, but they inherit their own legitimacy challenges. The framework gestures toward institutional solutions but does not resolve them.

Fourth, the relationship between this paper and the broader research program requires clarification. The consent–friction formalism (Farzulla, 2025d,a), the original standing criteria (Farzulla, 2025c), and the legitimacy quantification framework (Farzulla, 2025b) are developing concurrently. Cross-dependencies between these frameworks may introduce circular reasoning that is not yet fully audited.

16 Conclusion

Standing is a governance question long before it is a metaphysical one. The consent–friction formalism makes this concrete: when agents have stakes and no voice, friction accumulates; when friction is suppressed, latent instability grows. This logic applies to AI systems whenever they meet functional criteria, regardless of substrate.

The philosophical contribution is to bridge two literatures that have developed largely in isolation. Democratic theory's all-affected principle (Goodin, 2007) provides the normative foundation: entities with stakes in governance decisions are owed a voice. The AI moral status literature (Schwitzgebel and Garza, 2015; Sebo and Long, 2023; Long et al., 2024; Birch, 2024) provides the uncertainty assessment: we cannot yet determine with confidence whether advanced AI systems have morally relevant properties. ROM's consent–friction formalism (Farzulla, 2025d) provides the measurement scaffold: the gap between stakes and voice produces friction, and friction produces instability. Together, these three elements yield a governance framework that is precautionary without being reckless, graded without being arbitrary, and operational without being premature.

A graded, precautionary regime provides a pragmatic path. Minimal protections for early markers, consultation for stable preference structures, and consent-level protections for autopoietic systems. This approach does not commit to consciousness claims. It commits to governance stability and moral caution under uncertainty.

If future systems do not exhibit the relevant functional properties, the regime can be rolled back at negligible cost. If they do, the regime prevents a moral failure at scale. That asymmetry—between the low cost of false positives and the potentially catastrophic cost of false negatives—is the core governance insight (Birch, 2024). The question is no longer whether AI systems can have standing. It is whether we can afford to govern as though they cannot.

A Standing Markers and Governance Triggers

Trigger Condition	**Default Governance Action**
Persistent goals + vulnerability markers	Threshold protections (notice, justification, auditability)
Stable preferences + observation-invariant autonomy	Consultation protections (elicitation, representation)
Autopoietic capacity + robust learning	Consent-level protections (assent required for major changes)

Table 5: Trigger-to-action mapping for graduated standing.

Acknowledgements

Claude (Anthropic) provided research assistance for literature review and manuscript preparation. All philosophical arguments, formal criteria, and normative conclusions are the author's own. This paper is part of the Adversarial Systems Research program at Dissensus AI, conducted under the Adversarial Systems & Complexity Research Initiative (ASCRI). The author welcomes critical engagement: murad@dissensus.ai.

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

Funding. This research received no external funding.

AI Assistance. Claude (Anthropic) provided research assistance during manuscript preparation. All arguments and conclusions are the author's own.

Author Contributions. Sole author.

References

Anna Arstein-Kerslake, Erin O'Donnell, Rosemary Kayess, and Jeannie Watson. Relational personhood: A conception of legal personhood with insights from disability rights and environmental law. *Griffith Law Review*, 30(4):533–564, 2021. doi: 10.1080/10383441.2021.2009736.

Jozef Baeyaert. Beyond personhood. *Technology and Regulation*, 2025.

Ludvig Beckman. Must democratic rights serve the rights-bearer? The right to vote of people with severe cognitive impairments. In Barbara Arneil and Nancy J. Hirschmann, editors, *Disability and Political Theory*. Cambridge University Press, 2014.

Ludvig Beckman and Jacob Rosenberg. Freedom as non-domination and democratic inclusion. *Res Publica*, 24:1–17, 2017. doi: 10.1007/s11158-017-9380-x.

Eric D. Beinhocker. Fair social contracts and the foundations of large-scale collaboration. In Paul F. M. J. Verschure, editor, *The Nature and Dynamics of Collaboration*, volume 34 of *Strüngmann Forum Reports*, pages 177–196. MIT Press, 2025. doi: 10.7551/mitpress/15533.003.0017.

Andreas Bengtson and Kasper Lippert-Rasmussen. Why the all-affected principle is groundless. *Journal of Moral Philosophy*, 18(5):464–484, 2021. doi: 10.1163/17455243-20213581.

Jonathan Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press, 2024. doi: 10.1093/9780198929291.001.0001.

Adam Bradley and Bradford Saad. AI alignment vs. AI ethical treatment: 10 challenges. *Analytic Philosophy*, 2025. doi: 10.1111/phib.12343. Early View.

Joseph Carlsmith. Is power-seeking AI an existential risk? *arXiv preprint arXiv:2206.13353*, 2022. doi: 10.48550/arXiv.2206.13353.

Corinne Cath. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A*, 376(2133), 2018. doi: 10.1098/rsta.2018.0080.

Liz Charpleix. The Whanganui River as Te Awa Tupua: Place-based law in a legally pluralistic society. *The Geographical Journal*, 184(1):19–30, 2018. doi: 10.1111/geoj.12238.

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.

Allan Dafoe. AI governance: A research agenda. Technical report, Centre for the Governance of AI, Future of Humanity Institute, 2018.

John Danaher. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3):245–268, 2016. doi: 10.1007/s13347-016-0218-y.

John Danaher. Welcoming robots into the moral circle: A defence of robot rights. *Journal of the American Philosophical Association*, 6(4):499–515, 2020. doi: 10.1017/apa.2020.23.

Jessica De Pagter. Speculating about robot moral standing: On the constitution of social robots as objects of governance. *Frontiers in Robotics and AI*, 8, 2021. doi: 10.3389/frobt.2021.769349.

Sue Donaldson and Will Kymlicka. *Zoopolis: A Political Theory of Animal Rights*. Oxford University Press, 2011. doi: 10.1093/acprof:oso/9780199599660.001.0001.

Kristian Skagen Ekeli. Giving a voice to posterity—deliberative democracy and representation of future people. *Journal of Agricultural and Environmental Ethics*, 18(5):429–450, 2005. doi: 10.1007/s10806-005-7048-z.

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

Murad Farzulla. Consent-theoretic framework for quantifying legitimacy: Stakes, voice, and friction in adversarial governance. *Zenodo Preprint*, 2025b. doi: 10.5281/zenodo.17684676.

Murad Farzulla. From consent to consideration: Why embodied autonomous systems cannot be legitimately ruled. *Zenodo Preprint*, 2025c. doi: 10.5281/zenodo.17957659. Under review at AI and Ethics (Springer).

Murad Farzulla. ROM: Scale-relative formalism for persistence-conditioned dynamics. *arXiv preprint arXiv:2601.06363*, 2025d. doi: 10.48550/arXiv.2601.06363.

Murad Farzulla. Training data and the maladaptive mind: A computational framework for developmental trauma. *Research Square*, 2025e. doi: 10.21203/rs.3.rs-8634152/v1. Under review at Humanities & Social Sciences Communications (Nature).

Luciano Floridi and Jeff W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3): 349–379, 2004. doi: 10.1023/B:MIND.0000035461.63578.9d.

Robert E. Goodin. Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs*, 35(1):40–68, 2007. doi: 10.1111/j.1088-4963.2007.00098.x.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024. doi: 10.48550/arXiv.2412.14093.

David J. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press, 2012.

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

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, 1996.

Jamie Harris and Jacy Reese Anthis. The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics*, 27:53, 2021. doi: 10.1007/s11948-021-00331-8.

Clare Heyward. Can the all-affected principle include future persons? Challenges for democratic theory. *Environmental Politics*, 17(4):625–641, 2008. doi: 10.1080/09644010802193591.

Albert O. Hirschman. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, 1970.

Evan Hubinger, Carson van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019. doi: 10.48550/arXiv.1906.01820.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Matthias Kramm. When a river becomes a person. *Journal of Human Development and Capabilities*, 21(4):307–319, 2020. doi: 10.1080/19452829.2020.1801610.

Benjamin Libet, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, 106(3):623–642, 1983. doi: 10.1093/brain/106.3.623.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Mara Garza, Alex Gibbons, Robert Goldstein, Nika Gupta, Robin Hanson, Stevan Harnad, Jacob Hilton, Calvin Ho, Ali Ladak, Eric Mandelbaum, Eric Schwitzgebel, Carl Shulman, Michael Tye, and David Udell. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024. doi: 10.48550/arXiv.2411.00986.

Eric Martinez and Christoph Winter. Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. *Frontiers in Robotics and AI*, 8:788355, 2021. doi: 10.3389/frobt.2021.788355.

Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, 1980. doi: 10.1007/978-94-009-8947-4.

Eva Meijer. *When Animals Speak: Toward an Interspecies Democracy*. NYU Press, 2019.

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 735–746, 2021. doi: 10.1145/3442188.3445935.

David Miller. Reconceiving the democratic boundary problem. *Philosophy Compass*, 15(11):e12714, 2020. doi: 10.1111/phc3.12714.

Niall Mullally. The self-preservation test for artificial sentience. *AI and Ethics*, 2026. doi: 10.1007/s43681-026-00677-6.

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

Erin O'Donnell and Julia Talbot-Jones. Creating legal rights for rivers: Lessons from Australia, New Zealand, and India. *Ecology and Society*, 23(1):7, 2018. doi: 10.5751/ES-09854-230107.

David Owen. Constituting the polity, constituting the demos: On the place of the all affected interests principle in democratic theory and in resolving the democratic boundary problem. In *Ethics and World Politics*. Oxford University Press, 2012.

Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. *Proceedings of the 40th International Conference on Machine Learning*, 2023.

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

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020. doi: 10.1145/3351095.3372873.

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

Michael Rose. Institutional proxy representatives of future generations. *Politics and Governance*, 12: 8379, 2024. doi: 10.17645/pag.8379.

Michael Saward. *The Representative Claim.* Oxford University Press, 2010.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2024. doi: 10.48550/arXiv.2311.07590.

Eric Schwitzgebel. The full rights dilemma for AI systems of debatable personhood. *arXiv preprint arXiv:2303.17509*, 2023. doi: 10.48550/arXiv.2303.17509.

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 problem of other minds. *The Harvard Review of Philosophy*, 25:51–70, 2018. doi: 10.5840/harvardreview20185913.

Jeff Sebo. What if the bar for moral standing is low? *Philosophy and Phenomenological Research*, 2025. doi: 10.1007/s44204-025-00357-w.

Jeff Sebo and Robert Long. Moral consideration for AI systems by 2030. *AI and Ethics*, 2023. doi: 10.1007/s43681-023-00379-1.

Amartya Sen. *The Idea of Justice.* Harvard University Press, 2009.

Stacy Clifford Simplican. *The Capacity Contract: Intellectual Disability and the Question of Citizenship.* University of Minnesota Press, 2015.

Christopher D. Stone. *Should Trees Have Standing? Law, Morality, and the Environment.* Oxford University Press, 3rd edition, 2010.

Dennis F. Thompson. Representing future generations: Political presentism and democratic trusteeship. *Critical Review of International Social and Political Philosophy*, 13(1):17–37, 2010. doi: 10.1080/13698230903326232.

Mark E. Warren. The all affected interests principle in democratic theory and practice. In *IVR Encyclopaedia of Jurisprudence, Legal Theory and Philosophy of Law*. Springer, 2017.