PREPRINT v1.0.0 | Not peer-reviewed

Stakes Without Voice

A Governance Framework for AI Standing

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

[1] *Farzulla Research*

January 2026

Correspondence: murad@farzulla.org

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

Publication Metadata

DOI: 10.5281/zenodo.18195279
Version: 1.0.0
Date: January 2026

Note on Prior Work

This paper is version 1.0.0. It follows *From Consent to Consideration*, which introduced a functional criteria framework for political standing and made a conservative empirical claim that current systems do not meet those criteria. The present follow-up retains the criteria while integrating the ROM consent–friction formalism as a measurement scaffold and updating the empirical posture to a graded, precautionary view. It also expands the governance implications and clarifies safeguards against manipulation and liability externalization.

Research Context

This work continues the Adversarial Systems Research program, which investigates how stakes, voice, alignment, and information loss generate friction in multi-agent systems. The consent–friction formalism provides a cross-domain lens for political legitimacy, financial instability, and AI governance. The present contribution extends that framework to agentic AI by treating standing as a governance question: where stakes are real, voice exclusion is destabilizing regardless of substrate.

Acknowledgements

The author acknowledges the role of AI-assisted research tools for literature discovery and drafting support. All claims, theoretical structures, and normative arguments remain the author's own.

Methodologies: Research methodologies documented at `farzulla.org/methodologies`.

1 Introduction: From Criteria to Governance

The original *From Consent to Consideration* 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 follow-up makes three moves. First, it connects the criteria to ROM's consent–friction formalism, 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.

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.

The criteria proposed in the original paper 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.

3 Methods and Scope

This paper is normative and theoretical. It synthesizes political legitimacy theory, AI ethics scholarship, and the ROM consent–friction formalism 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.

4 The Consent–Friction Scaffold (ROM)

The Replication-Optimization Mechanism (ROM) 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.

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.

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. 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. 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. This makes standing a governance problem without assuming metaphysical consensus. 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-style experiments suggest 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. 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. 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.

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.

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. Highly capable systems might engage in scheming—strategic planning to achieve goals despite oversight.

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 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. Capability control addresses symptoms rather than causes, potentially intensifying friction by increasing stakes while maintaining zero voice.

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.

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 standing question is whether such behavior indicates stakes that governance

must take seriously, not whether the system is "alive" or conscious.

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.

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. 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 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 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

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.

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.

- **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, 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.

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.

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.

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 Study: 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.

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 point is not to protect the system for its own sake but to prevent unaccountable harm. The standing proxy operates as a governance lever.

This example matters for AI & Society because it links 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. The standing framework provides a route to formalize that accountability without resorting to metaphysical personhood.

11 AI & Society Positioning: Societal Embedding and Public Accountability

AI & Society has always been about the entanglement between systems and institutions. Agentic AI intensifies that entanglement by inserting systems into decision loops previously reserved for human or organizational actors. 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.

This paper's contribution to the AI & Society agenda 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.

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 cul-de-sac 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.

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.

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, and future generations. The philosophical precedent for partial standing is robust.

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. The framework also assumes that friction is a relevant governance objective; agents that actively prefer instability fall outside the bridge principle. These limitations are not fatal, but they are real.

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.

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. If they do, the regime prevents a moral failure at scale. That asymmetry is the core governance insight.

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.