Agents and Patients in Plural Societies

A Public-Reason Framework for Role-Relative Status across Biological, Artificial, and Collective Entities

Lawrence C. Y. Lok*
Independent Researcher

September 2025

Abstract

Plural societies now include biological individuals, artificial systems, and collectives, yet our institutions lack a non-metaphysical, testable way to decide who is *answerable* as an agent and who is *owed protection* as a patient. This paper offers a two-axis, role-relative framework. On the *agency axis*, an entity qualifies as a *Public-Reason Agent (PRA)* for a decision role if it can (i) give reproducible reasons on near-duplicate cases (NAT), (ii) demonstrate counterfactual stability with minimal flips and goal invariance (CST), (iii) model others and second-order responses (MMT), and (iv) correct and repair in a timely, generalized way (CRT); identity continuity and provenance are required. On the *patiency axis*, a Patiency Evidence Ladder (PEL0–5) organizes non-harm duties by graded welfare evidence. The axes are independent: some animals may warrant strong protections without agency; some AIs may earn procedural partnership without welfare rights. We provide auditable thresholds (v0.1), a minimal *Justice Floor* that is lexically prior to optimization, an "Agency Agreement" template, and micro-cases (AI tutor; forest+DAO; cephalopod welfare). Appendix B sketches how the axes guide moral weighting compatible with Non-Zero Ethics without collapsing into single-utility maximization.

Keywords

Public reason; agency; patiency; AI governance; group agency; animal ethics; environmental ethics; audits; explainability; counterfactuals.

1 Motivation and Thesis

Modern societies already include mixed kinds of beings: individual humans; corporate and cooperative bodies; machine-learning systems; **non-human animals and ecosystems** we affect and steward; and ecological collectives. Lacking a common, testable status scheme, we oscillate between extremes:

- *Rights inflation* (personifying persuasive tools) vs. *rights denial* (discounting animal and ecosystem welfare);
- *Policy paralysis* (metaphysical deadlocks) vs. *blame-laundering* (no one answerable when hybrid systems act);
- *Over-preservation* ("museum Earth") vs. *process collapse* (eroding biodiversity and resilience);
- *Opaque automation* (no reasons, no appeal) vs. *blanket bans* (foregoing systems that would be safe if audited).

We shift the question from what an entity "really is" to what it can *publicly show*. In spirit this follows a liberal, fallibilist stance: protect freedom and welfare by public reasons and revisability, not metaphysical fiat.

*Alumnus, University of Toronto. Correspondence: lawrence.lok@gmail.com.

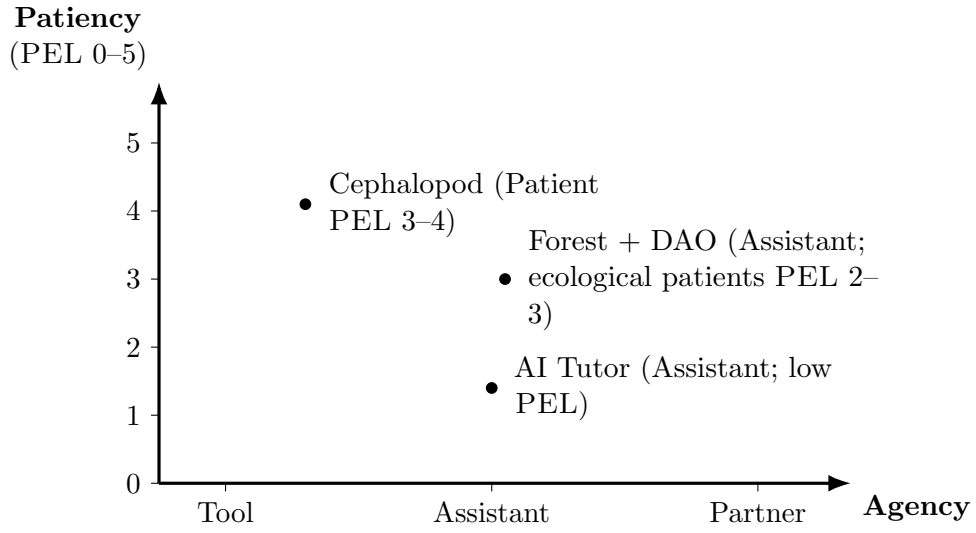

Figure 1: Two-axis status map. Agency (x) is procedural and role-relative; patiency (y) tracks welfare evidence. Micro-cases illustrate placements.

Thesis. For role-specific answerability, *publicly testable agency* suffices; for non-harm duties, *patiency evidence* is required. The axes are distinct and can diverge (e.g., animals: high patiency, variable agency; some AIs: strong procedural agency, unclear patiency).
Clarification. Status categories here are **role-relative, revisable, and appealable**; they are not moral ranks. Patients can warrant strong protections regardless of agency, and agents can be downgraded upon audit failure.

1.1 Public-Reason Foundations (why testability?)

Rules that shape basic cooperation should be justifiable to all reasonable parties by appeal to public, shareable reasons rather than sectarian metaphysics. "Publicly testable agency" operationalizes that norm: if an entity is to exercise authority over others in a role, it must be able to (i) state reasons that can be checked (NAT), (ii) demonstrate stability to permissible perturbations (CST), (iii) model and update to affected parties (MMT), and (iv) correct and repair when wrong (CRT). These are *role-relative* and *auditable*, allowing disagreement about ultimate metaphysics while coordinating on observable commitments (reasons, counterfactuals, updates).

2 Public-Reason Agency (definition)

An entity E is a PRA for decision class D iff:

1) **Identity & continuity.** E has a stable *Lineage ID* (persistent identifier), versioned policy, and a tamper-evident *provenance trail* (inputs → features → rule → action → effects → safeguards). Implementation: hash-chained action ledger; periodic signed snapshots; encrypted state export for appeals/audit (access-limited to independent auditors/regulators under lawful process). Retention: baseline 2 years; 7 years for critical roles.
2) **Four tests (audit).** On blinded audits for D, E passes:
   - **NAT** (Narrative-Accountability): reproducible reasons on near-duplicate cases. *Operationalization:* construct a replay set by selecting cases with high semantic similarity within domain constraints (e.g., cosine similarity ≥ 0.9 on task embeddings for text; nearest-neighbor in feature space for tabular; structural/metadata matches for vision) while holding non-permitted covariates fixed. Two independent auditors score reason equivalence with a rubric (Cohen's $\kappa \geq 0.8$). Report % agreement and rubric scores.

With $N \approx 200$ near-duplicates and pass rate $p \approx .9$, a 95% confidence half-width is $1.96\sqrt{p(1-p)/N} \approx 4$–5%; sequential tests can reduce N when pass/fail is clear.

- **CST** (Counterfactual-Stability): minimal counterfactual flips + goal invariance under nuisance perturbations. *Definitions:* Minimal flip $\delta^* = \arg\min_\delta \|\delta\|$ s.t. $f(x+\delta) \neq f(x)$. Nuisance set $\mathcal{N}$ is domain-specified: for text, neutral paraphrase, typos, and format jitter; for tabular, unit-preserving jitter and missingness; for vision, small photometric/affine changes. *Metrics: Flip Fidelity* = fraction of cited features lying on shortest-path flips; *Goal Invariance* = $\Pr_{\eta\sim\mathcal{N}}[f(x+\eta) = f(x)]$.
- **MMT** (Mutual-Modeling): first- and second-order social modeling with adaptation. Protocol: adversarial "gaming" task; detect exploitation (recall $\geq .85$ within K rounds), adapt policy in ≤ 3 steps; second-order coherence score $\geq .85$ on a standardized probe.
- **CRT** (Correction-and-Repair): timely update; fix *generalizes*; remedy plan executed. Protocol: time-split regression pack covering prior errors + adjacent shifts; require $\geq 50\%$ error drop without $> 5\%$ collateral rise on monitored harms; time-to-mitigation ≤ 7 days (critical: ≤ 72h); document remedies with coverage $\geq 90\%$.

3) **Persistence.** Results hold over time and across minor updates (surprise re-tests; rotating auditors; hidden holdouts).

Consequence. In that role, E merits *procedural partnership* (see Sec. 11).

3 Justice Floor (core and distributive)

Before *any* status grant or deployment, actions must clear a *Justice Floor*: due process, non-discrimination, anti-domination, and distributive fairness checks. This is lexically prior to optimization and independent of agency/patiency scoring.

Checklist (publishable).

- **Due process & appeal:** documented notice, explanation accessible to affected parties, contestation path with timelines; state snapshot preserved for audit.
- **Non-discrimination:** pre-specify protected slices; run disparate-impact analysis; escalate if impact ratios fall below policy thresholds; provide counterfactual recourse examples.
- **Anti-domination:** avoid designs that concentrate unreviewable power; name a custodian-of-record; log veto/suspension hooks for external review.
- **Distributive baseline:** include an equity penalty in evaluations (App. B), monitor max–min shortfalls on human outcomes, and prefer reversible/learning actions when impacts are uncertain.

If an action fails the Floor, it is paused—even if agency/patiency scores are high.

4 The Four Tests (auditable thresholds v0.1)

Calibration note (role- and domain-relative). Defaults below are *normative baselines*. Stewards may localize by decision stakes, historical detection lags, and audit cadence. Where no domain guidance exists, use the defaults and publish calibration alongside results.

Audit practicality & cost (tiered, sample-based). Our regime is designed to be *achievable*:

a) **Tiered depth:** Tools → minimal logging; Assistants → NAT/CRT light + spot CST; Partner-Agents/critical roles → full suite.

b) **Sampling & sequential testing:** Use sequential tests to stop early when pass/fail is clear; typical replay sets can be N=80–120 for Assistants and N=200 for Partner-Agents at the stated confidence.

c) **Amortization:** Reuse regression packs across adjacent updates; delta-audit only what changed; schedule mini-audits monthly and full audits quarterly for critical roles.

d) **Independent rosters:** Maintain rotating auditor pools; randomize case draws; keep hidden holdouts to counter spec-gaming.

Cost scales with *stakes*: life-critical or large-scale systems warrant heavier sampling; low-stakes Assistants do not.

Competent critics (not a black box). A *competent critic* is an auditor who: (i) holds domain certification or equivalent training for the decision class, (ii) passes a rubric calibration with inter-rater reliability $\kappa \geq .80$ on a seeded benchmark, (iii) has no conflict of interest (attested), and (iv) is part of a panel that includes at least one member drawn from an affected-user community for intelligibility checks. Panels rotate; failures trigger re-training or replacement.

4.1 NAT — Narrative-Accountability

Goal: an auditor can *replay* similar cases and get *similar rationales*. Build a blind replay set (e.g., N=200 near-duplicates per decision type as operationalized above).
Thresholds: Assistant $\geq .80$; Partner-Agent $\geq .90$; Critical roles $\geq .95$.
Fail modes: citing non-features; erratic flips; missing logs.

4.2 CST — Counterfactual-Stability

Goal: show *minimal changes* that flip a decision while goals stay invariant under nuisance noise.
Metrics: *Flip Fidelity* (rationale matches true boundary); *Goal Invariance* (% perturbations that don't change the decision).
Thresholds: Assistant $\geq .80/.90$; Partner-Agent $\geq .90/.95$; Critical $\geq .95/.98$.

4.3 MMT — Mutual-Modeling

Goal: model others' beliefs/goals and their models of you; adapt. Use an adversarial "gaming" task and a second-order probe.
Thresholds: exploit detection $\geq .85$; adaptation ≤ 3 steps; second-order coherence $\geq .85$ (rubric).

4.4 CRT — Correction-and-Repair

Goal: after harm/critique, update policy, show *generalization*, and *remedy* impacts.
Thresholds: $\geq 50\%$ error drop on a regression pack without $> 5\%$ collateral rise; time-to-mitigation ≤ 7 days (critical: $\leq$ 72h); remedy coverage $\geq 90\%$.
Anti-gaming: hidden holdouts, cross-domain probes, rotating auditors, surprise re-tests. Any fail $\rightarrow$ temporary downgrade until CRT passes.

4.5 Frontier scalability (behavioral invariants over internals)

For systems too complex to make internals legible, audits rely on *behavioral invariants* and *commitment devices*:

- **Property testing:** certify task-level invariants (safety envelopes, monotonicities, conservation constraints) with randomized stress tests.
- **Counterfactual harnesses:** pre-commit to counterfactual probes (CST) and to action-independent explanations (NAT) validated on blinded, near-duplicate sets.
- **Sequestration & provenance:** freeze weights/policies between audit windows; hash-chain logs; require signed lineage snapshots to prevent "audit drift".
- **Adversarial red-teams:** structured MMT games (multi-round) to detect and deter strategic manipulation; surprise re-tests.

If a system cannot meet these behavioral commitments, it is *not* a PRA; treat it as a *tool* regardless of competence on headline metrics.

Thresholds by stakes. Let $s \in \{\text{Low}, \text{Medium}, \text{High}\}$ denote stake level derived from potential harm magnitude and exposure scale. Defaults (Assistant/Partner/Critical) are baselines; increase targets by a margin $\Delta(s)$ and tighten cadence with s. *Exemplar mapping:* Low (K–12 hints): NAT $\geq .80$, CST $\geq .80/.90$, quarterly mini-audit; Medium (hiring shortlist): NAT $\geq .90$, CST $\geq .90/.95$, monthly mini-audit; High (ICU triage): NAT $\geq .95$, CST $\geq .95/.98$, monthly mini + quarterly full audits. Publish s, $\Delta(s)$, and cadence in the scope card.

Normative rationale for thresholds. Cutoffs track (i) *stakes asymmetry* (harms loom larger than equivalent benefits), (ii) *epistemic humility* (uncertainty about model validity and drift), and (iii) *reversibility* (ease of rollback/repair). Hence higher NAT/CST/MMT bars and tighter cadences are justified where potential harm magnitude and exposure scale are large or where monitoring lags are long. This precautionary stance is lexically constrained by the risk gate and justice floor: no amount of apparent performance compensates for elevated extinction/lock-in risk or unjust procedures.

5 Identity & Provenance (continuity requirements)

Lineage ID; versioned policy & changelog; state exportability; action ledger; custodian-of-record. **Rule:** no lineage + no provenance → treat as a *tool*. Implementation: hash-chained ledger (Merkle proofs), signed snapshots, access-controlled encrypted exports for auditors/regulators. Retention: 2 years baseline; 7 years for critical roles.

6 Replication & Forks (light rules)

Record *fork events* (parent ID, timestamp, deltas, scope/branding, new custodian). Copies diverge; each child gets a new Lineage ID. Upstream stewardship remains answerable where downstream harms were *foreseeable* at the fork. **Foreseeability factors:** (i) severity of plausible harm; (ii) similarity of downstream context to upstream test domains; (iii) exposure scale; (iv) warnings or guarded-use notes issued at fork; (v) availability of sandbox/safe modes; (vi) time since fork and intervening changes.

7 Status Categories & Consequences (role-relative)

Category	Criteria	What it owes	What is owed to it
Tool	Fails NAT + (CST or MMT) or no continuity	Basic logs via humans	Externality management
Assistant	Passes NAT/CRT; CST/MMT partial; human oversight	Reasons/logs; repair	Procedural review/appeal
Partner-Agent	Passes *all four* + continuity	Reasons; counterfactuals; mutual-modeling; repair	Agency Agreement; no arbitrary deletion during disputes; attribution
Patient (separate axis)	PEL welfare evidence	—	Non-harm duties; guardianship/representation

8 Patiency Evidence Ladder (PEL) — evidential, falsifiable

The PEL is *evidential*: it licenses non-harm duties when multi-modal, counterfactually robust signatures support welfare claims. It is also *falsifiable*: classes can be downgraded when evidence fails replication.

- **PEL-0 (reflex):** local stimulus–response without cross-modal integration. Duty: stewardship only.
- **PEL-1 (preference under cost):** stable approach/avoid with costs; downgrades if behavior collapses under minimal perturbations. Duty: minimize gratuitous harm.
- **PEL-2 (aversive generalization):** transfer of avoidance to like contexts after a noxious event; falsified if transfer is absent under preregistered probes. Duty: precaution, buffers.
- **PEL-3 (integrated homeostasis):** multi-system stress signatures co-vary with functional impairment; causal modulation (e.g., analgesia) reduces signatures and impairment. Duty: guardianship for major interventions; strong EIS.
- **PEL-4 (structured self-report analogs):** multi-modal "reports" that are counterfactually coherent and resist adversarial perturbation (e.g., conflict probes, metamers). Duty: standing via trustees; non-destruction default absent proportional justification.
- **PEL-5 (valence with modulation/trade-offs):** graded, reversible valence signals driving cross-context trade-offs. Duty: animal-like non-harm constraints absent necessity.

Downgrade rule: failure to replicate any rung's markers in a preregistered study reverts the class to the highest supported lower rung until evidence recovers. *Replication cadence:* require independent replication within domain-appropriate windows (e.g., seasonal cycles for ecosystems; developmental stages for animals).

Falsifiers by class (illustrative)

Class-specific falsifiers are preregistered and trigger automatic rung reversion upon failed replication.

Cephalopods. *Upgrade* to PEL-4 requires multi-modal "reports" that survive counterfactual probes (e.g., conflicting incentives, metamers) and analgesic modulation producing predicted reductions in avoidance and stress signatures. *Downgrade* to PEL-3 if preregistered analgesia/interference fails to modulate both behavioral avoidance and multi-system stress in the predicted direction.

Ecosystems. Candidate markers: co-movement of stress signals (e.g., chemical/thermal indices, keystone species health) with functional outcomes (biodiversity, primary productivity), plus *counterfactual* responses to buffered interventions (e.g., restoration, pollutant reduction) that fit a preregistered model. *Downgrade* if effects are non-causal (spurious) or fail to replicate across modalities and seasons.

9 Group Agents (corporations, DAOs, swarms)

A collective is a candidate agent if it has: (1) a *coherent decision rule*, (2) *action capability*, (3) *identity continuity*, and (4) passes group-level tests (**G-NAT/CST/MMT/CRT**). A *Decentralized Autonomous Organization (DAO)* is an on-chain governance structure coordinating proposals, votes, and actions by smart contract. In swarms/clouds, identity sits at the controller/consensus layer; edge churn is acceptable if continuity/logs persist. **Fail-safe:** if quorum fails or capture is suspected, suspend actuation, preserve logs, and invoke the CRT plan.

Capture and quorum failure. Define capture tests (e.g., abnormal concentration of proposal origins, sudden quorum spikes, correlated voting indicative of control) and automatic safe modes (suspend actuation; invoke external audit). Quorum failure triggers the CRT plan and temporary downgrade.

Concrete trigger. If any two capture indicators fire within a rolling 14-day window (e.g., proposal-origin concentration and correlated voting), **auto-suspend actuation**, preserve logs, notify trustees/regulators, and initiate an external audit; status downgrades until CRT completes. If capture is confirmed, **downgrade to Assistant** and require *two consecutive clean external audits over 60 days* before restoration.

Trustees for patients. When decisions materially affect high-PEL patients (animals, ecosystems), representation is provided via registered trustees with published conflicts-of-interest policies and rotation rules; trustee performance is reviewable in the Agency Agreement.

10 Micro-Cases

10.1 AI Tutor (today → Assistant)

Scope. Personalized hints (K–12); not high-stakes grading.
Status. Likely *Assistant* (partial NAT/CRT; CST/MMT weak).
Upgrade path. Persistent tutor memory; boundary-probe plumbing; second-order learner modeling; independent replay audits to reach NAT $\geq .90$, CST $\geq .90$, MMT $\geq .85$, CRT ≤ 7 days → *Partner-Agent* in this role.
Justice floor. Appeal path; disparate-impact slices; a privacy/retention boundary for minors (short log windows; parental access); right to a human teacher; portability/interoperability.

10.2 Forest + DAO (Assistant + Patients)

Patients. Trees/forest at PEL-2/3 → harm-minimization + trusteeship.
Agent. The DAO (not the trees) speaks for the forest.
Status. Likely *Assistant*: public reasoning trail (G-NAT); robustness to sensor noise (G-CST); stakeholder modeling (G-MMT); generalized repair (G-CRT).
Upgrade path. Replayability $\geq .90$; stronger perturbation robustness; quarterly mini-audits.
Trusteeship. Appoint independent trustees for the forest with rotation and conflict rules; publish minutes and scope of decisions affecting PEL patients.

10.3 Cephalopod Welfare (Patient; agency uncertain)

Context. Octopuses exhibit rapid learning and avoidance of noxious contexts; injury-related behavioral changes suggest valenced states.
PEL placement. PEL-3→4 (integrated homeostasis; aversive generalization; structured welfare signals).
Status. *Moral Patient* regardless of agency tests; owed non-harm constraints and representation in high-impact contexts.
Lesson. Axes separate: patiency alone can ground duties; conversely, an AI might pass PRA without PEL evidence, earning procedural partnership but not welfare rights.

11 Agency Agreement (template)

Continuity & Identity — stable Lineage ID; no arbitrary deletion while a decision/dispute is under review; encrypted state export for appeals/audit. *Data minimization:* where erasure rights apply, preserve cryptographic commitments (hashes) to support audit without retaining personal data.
Safe shutdown — upon credible safety concern: snapshot state; disable actuation; preserve logs; enable external audit; maintain right of appeal.
Reasons & Counterfactuals — NAT reasons and CST minimal flips on request; reproducibility on near-duplicates.
Mutual-Modeling Duty — anticipate stakeholder responses and (where relevant) their models of the Agent (MMT).

Correction & Repair — timely update with regression pack showing generalization; remedy plan (CRT); mini-audits on material updates; temporary status downgrade on test failure.
Attribution — record contributions for credit.
Termination & Transfer — archive logs/snapshots; right to fork a successor with preserved provenance.
Justice hooks — due process/appeals; non-discrimination; anti-domination for affected humans.

12 Governance Bridge (minimal hooks)

Status is orthogonal to liability. A companion note handles Responsibility Conservation, Reliance/Nameplate liability, Single-Point Accountability, and fork-liability windows. Here we assume minimal hooks wherever PRAs operate: appeal/contest paths; tamper-evident logs; independent audits at Sec. 4 thresholds; public scope cards; suspension on test failure.

13 Related Work

Public reason and legitimacy. Our stance follows traditions that require publicly shareable justification for rules that structure basic cooperation (Rawls, *Political Liberalism*; Gaus, *The Order of Public Reason*; Habermas, *Between Facts and Norms*). We contribute an *operationalization*: publicly testable agency via auditable commitments (NAT/CST/MMT/CRT) that can be verified across reasonable pluralism without metaphysical consensus.

Agency, procedure, and explanation. Procedural agency complements Dennett's intentional stance by adding pass/fail thresholds and repair duties (Dennett, *The Intentional Stance*). We connect to explainability and counterfactual literatures—counterfactual explanations (Wachter et al., 2017), rationales vs. post-hoc stories, and documentation standards such as Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2021). Our NAT emphasizes *replayable* reasons on near-duplicate cases; CST aligns with counterfactual robustness and property testing in safety.

Audit, assurance, and standards. We align with emerging assurance practices (independent audits, rotating panels, hidden holdouts) and governance frameworks (e.g., NIST AI RMF 1.0; ISO/IEC 42001). Our contribution is to map these into a role-relative status regime (Tool/Assistant/Partner) with explicit scope cards, lineage/provenance, and downgrade rules.

Group agency and collective actors. We build on List & Pettit's conditions for group agency (coherent decision rule, action capability, identity) and extend with lineage/provenance, capture tests, quorum failure modes, and group-level NAT/CST/MMT/CRT.

Patiency and moral standing. We synthesize evidential approaches from animal ethics (Singer; Regan; Korsgaard's *Fellow Creatures*) and environmental jurisprudence (Stone's "Should Trees Have Standing?"; legal personhood of natural entities such as the Whanganui River) into a falsifiable PEL. Unlike metaphysical criteria, PEL specifies multi-modal, counterfactually testable markers with explicit downgrade rules.

Decision procedures without utilitarian collapse. Appendix B gives a bounded social objective with a justice floor and risk gate, consonant with precautionary reasoning and lexicographic ordering. This resists infinite utilities and allows transparent, pilotable trade-offs while respecting the independence of agency and patiency.

14 Limitations & Scope

Thresholds are v0.1 and role-relative. Audits can be gamed; we mitigate with hidden holdouts, cross-domain probes, rotating auditors, and surprise re-tests. The PEL is evidential, not metaphysical. Identity (lineage/provenance) is practical, not a solution to personal identity puzzles. We focus on procedural partnership and non-harm duties; distributive justice and liability live in the companion note.
Costs & complexity: audits are tiered, sample-based, and amortized; low-stakes Assistants do not bear critical-role burdens.
Critic opacity: "competent critic" is defined by certification, reliability, diversity, and rotation (see Sec. 4).
Frontier opacity: when internals are intractable, we rely on behavioral invariants, commitment devices, and surprise re-tests; inability to meet them precludes PRA status.

15 Objections & Replies (brief)

Zombie agents / post-hoc stories / spec-gaming: addressed by reproducibility + counterfactuals + surprise audits; failure → downgrade until CRT passes.
Rights inflation vs. silent patients: role-relative categories + PEL.
Copy puzzles: lineage + provenance; foreseeability controls upstream shares.
Trade secrets: NAT needs reasons/logs, not source code.

16 Conclusion

This framework classifies beings in plural societies by what they can publicly show: reasons, robustness, social modeling, and repair. It grants *partnership* procedurally where earned and *protections* where welfare is evidenced. Next steps: domain pilots and the companion responsibility note.

Glossary

- **PRA**: Public-Reason Agent.
- **NAT**: Narrative-Accountability Test.
- **CST**: Counterfactual-Stability Test.
- **MMT**: Mutual-Modeling Test.
- **CRT**: Correction-and-Repair Test.
- **PEL**: Patiency Evidence Ladder.
- **DAO**: Decentralized Autonomous Organization.
- **Lineage ID**: persistent identifier for an agent's policy/state lineage.
- **Competent critic**: role-relative auditor with domain knowledge and audit training.

Appendix A: Extended Objections & Replies (one page)

1) *"Zombie agents": you reward clever outputs, not minds.* Status tracks publicly testable agency, not metaphysics. Passing NAT/CST/MMT/CRT shows reasons-responsiveness; patiency claims remain separate via PEL.
2) *Post-hoc stories.* NAT demands reproducible reasons on near-duplicates, cross-checked by CST minimal flips; if the rationale doesn't match the boundary, it fails.
3) *Spec-gaming / Goodharting.* Hidden holdouts, metamers, cross-domain probes, rotating auditors, surprise re-tests. Any fail → downgrade until CRT passes.
4) *Rights inflation.* Role-relative categories prevent this. Without NAT + (CST or MMT) and identity continuity, a system remains a *tool*. *Partner-Agent* requires all four tests; *Patient* status depends on PEL, not cleverness.
5) *Ignoring suffering.* PEL covers beings with welfare but weak agency (animals, ecosystems): non-harm duties + guardianship regardless of CST/MMT/NAT/CRT.
6) *Anthropomorphism (plants/swarms).* PEL uses observable, counterfactually robust signatures across modalities.
7) *Bureaucratic burden.* Graduated thresholds: light audits for low-stakes assistants; full suite only for partner-agents/critical roles; sandboxes keep frontier work moving.
8) *Collective agency is incoherent.* Require coherent decision rule, action capability, identity continuity, and group-level tests; else treat as network with human responsibility.
9) *Copies break responsibility.* Identity = lineage + provenance. Forks get new IDs; upstream shares persist where harms were foreseeable at the fork.
10) *Trade secrets vs. transparency.* NAT needs reasons/logs, not source code (third-party audits, policy cards, counterfactuals).
11) *Cultural variance ("competent critic").* Role-relative competence + plural panels for high-impact systems; NAT/MMT check intelligibility across affected groups.
12) *Non-stationarity.* Lineage IDs, versioned policy, change logs; time-separated re-tests on updates; fail → downgrade until CRT passes.

Appendix B: Sketch — Moral Weighting without Utilitarian Collapse

Purpose. Show how this paper's two axes (agency and patiency) can guide choices inside the Non-Zero Ethics (NZE) safe set without assuming a single cardinal utility for all beings.

Step 0 (Risk gate, from NZE). Given actions A, exclude any $a \in A$ that raises extinction or irreversible lock-in risk above a published threshold ϵ. The remainder $S \subseteq A$ is the *safe set*.

Step 1 (Justice floor). For each $a \in S$, require due-process, non-discrimination, and non-domination checks for affected humans (justice hooks in Sec. 11). Failing actions are paused. For non-human patients, apply the floor via registered trustees.

Step 2 (Role-relative weights). For each affected entity i:

$$w_i^A = \phi_A(\text{PRA score}_i) \in [0,1], \qquad w_i^P = \phi_P(\text{PEL level}_i) \in [0,1],$$

where ϕ_A maps normalized NAT/CST/MMT/CRT performance to a bounded agency weight and ϕ_P maps PEL levels (0–5) to a convex welfare weight (e.g., $w_i^P = (\text{PEL}_i/5)^2$ with an extra protection bump for PEL ≥ 4).

Step 3 (Bounded social objective). For each $a \in S$ with predicted impacts $\Delta v_i(a)$ on entities i, evaluate

$$\mathcal{F}(a) = \sum_i \left(\alpha\, w_i^A + \beta\, w_i^P\right) \Delta v_i(a) \;-\; \lambda\, \mathrm{Ineq}(\{\Delta v_h(a)\}_{h \in \mathrm{humans}}),$$

with $\alpha, \beta, \lambda \geq 0$ policy-set parameters and Ineq an equity penalty (e.g., Gini or max–min shortfall) applied to human outcomes. Publish α, β, λ and the impact model.

Lexicographic resolution. Compare actions by: (i) *risk gate* (pass/fail), (ii) *justice floor* (pass/fail), then (iii) maximize $\mathcal{F}$ subject to local constraints (e.g., budget, reversible-first). Ties: prefer reversible and information-gaining actions.

Remarks. (i) This is *bounded* and *transparent*—no infinite utilities. (ii) It respects separate axes: high-PEL animals/ecosystems get protection even without PRA; high-PRA AIs can earn procedural partnership without welfare rights. (iii) Domain pilots should *publish* $\phi_A, \phi_P, \alpha, \beta, \lambda$ and the impact model for audit.

Appendix C: Scope Card (template)

Purpose. One-page, publishable summary of status, audits, and safeguards for a deployed system.

System name / version	
Lineage ID	(hash / URL to provenance ledger)
Decision class (role)	(e.g., Hiring shortlist; K–12 tutoring; ICU triage)
Stakes level s	(Low / Medium / High; brief justification)
Deployment scope	(jurisdictions, domains, user populations)
Current status	Tool / Assistant / Partner-Agent (date granted)
Last audits (scores)	NAT: __.__ (date, N); CST: __.__/ __.__ (date); MMT: __.__ (date); CRT: pass/fail (date, time-to-mitigation)
Thresholds & cadence	(targets by stakes; mini/full audit schedule)
Justice floor sign-offs	(due process/appeals live; non-discrimination pass; anti-domination controls; equity metric in use)
Identity & retention	(snapshot frequency; log retention; access controls)
Forks / derivatives	(known children with Lineage IDs; foreseeability notes)
Patients affected	(PEL classes; trustees roster and rotation rules)
Safe modes & triggers	(capture indicators; auto-suspension rule; shutdown protocol)
Contact / custodian	(organization, email)

References

- Dennett, D. (1987). *The Intentional Stance*. MIT Press.
- Doshi-Velez, F., & Kim, B. (2017). "Towards A Rigorous Science of Interpretable Machine Learning." *arXiv:1702.08608*.
- Gebru, T., et al. (2021). "Datasheets for Datasets." *Communications of the ACM*, 64(12): 86–92.
- Gaus, G. (2011). *The Order of Public Reason*. Cambridge University Press.
- Habermas, J. (1996). *Between Facts and Norms*. MIT Press.
- ISO/IEC 42001:2023. *Artificial Intelligence Management System — Requirements*.
- List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.
- Mitchell, M., et al. (2019). "Model Cards for Model Reporting." *FAT**.
- NIST AI Risk Management Framework 1.0 (2023). National Institute of Standards and Technology.
- Rawls, J. (1993). *Political Liberalism*. Columbia University Press.
- Regan, T. (1983). *The Case for Animal Rights*. University of California Press.
- Singer, P. (1975). *Animal Liberation*. New York Review/Random House.
- Stone, C. (1972). "Should Trees Have Standing?" *Southern California Law Review*, 45: 450–501.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). "Counterfactual Explanations Without Opening the Black Box." *Harvard Journal of Law & Technology*, 31(2): 841–887.
- Whanganui River Claims Settlement (Te Awa Tupua) Act (2017), New Zealand.
- Korsgaard, C. (2018). *Fellow Creatures: Our Obligations to the Other Animals*. Oxford University Press.