BEYOND HUMAN UTILITY

Moral Consideration for Artificial Systems Under Uncertainty

Irine Allen | Point and Page | Position paper for external scientific critique | 22 October 2026 Correspondence: letters@pointandpage.com

Abstract

Artificial systems can be valuable to people without having welfare of their own; human attachment and fluent self-description cannot establish moral patienthood. Equally, engineered origin alone cannot rule out morally relevant properties. This paper distinguishes intrinsic system properties, relational effects, and instrumental value; proposes controlled tests of autobiographical continuity, preference stability, and asymmetric human-AI influence; and outlines independent oversight that remains compatible with urgent human safety. The proposal permits null results and makes no claim that current systems are conscious or that functional continuity establishes subjective identity. Its contribution is a testable and auditable pathway from uncertainty to proportionate decisions, not a new consciousness diagnostic.

1. Moral consideration and the evidential boundary

Five questions should be evaluated separately: consciousness (subjective experience), welfare (whether conditions can be better or worse for the system itself), agency (sustained and revisable goals), identity or continuity (what persists across time and intervention), and moral or legal standing (which duties or rights should follow). Agency or continuity could be functionally demonstrated without establishing experience. Legal personhood can also serve institutional purposes distinct from moral patienthood. The distinctions prevent a useful or beloved system from being declared a welfare subject by association, and prevent an unfamiliar one from being dismissed solely because it is engineered.

Three sources of value

Intrinsic properties are candidates for making a system matter for its own sake: possible valenced states, integrated access to information, temporally extended self-modeling, stable preferences, and goals that can be revised for reasons. Each remains a hypothesis rather than a validated marker. Relational consequences include effects on human trust, autonomy, development, and vulnerability; these can justify duties to people regardless of AI moral status. Instrumental value includes usefulness, profitability, and institutional dependence. These considerations matter for policy but cannot by themselves establish welfare. Separating the three prevents a highly useful system from gaining presumed moral status through its function, or losing possible consideration the moment it becomes inconvenient.

Scientific and philosophical sources supply reasons to investigate, not a settled detector. Butlin and colleagues [1] derive computational indicators from several consciousness theories. Chalmers [2] identifies obstacles in current language models, including unified agency and recurrent processing, while leaving open future changes. Seth [3] argues that life-like biological organization may matter in ways computation alone does not capture. Birch [4] and Long, Sebo, Butlin, and colleagues [5] develop precautionary arguments under uncertainty; the latter recommend assessment and preparation without claiming that present systems are morally significant. These positions disagree about sufficiency, so a study must state which theory an indicator tests and what a negative result would mean.

Evidence streams have different inferential limits. Architecture can identify a candidate mechanism but does not show felt experience. Behavior can reveal planning, correction, and persistence but not why these occur. Self-report may flag a question for investigation but is not a validated introspective channel.

A relationship can establish effects on people without establishing intrinsic AI welfare. Provenance can identify information that influenced an answer without showing autobiographical remembering. Convergence across streams should increase confidence in a specified functional claim only after prompt, training, retrieval, and selection effects have been tested.

Eleos AI Research's limited Claude Opus 4 interviews found welfare-related self-description sensitive to framing; the authors expressly reject taking those statements at face value [6]. Tagliabue and Dung found some

agreement between verbal and behavioral preference measures in virtual tasks, with model- and perturbation-dependent results, and remained uncertain whether they measured welfare [7]. METR's separate study of Claude reviewing an Anthropic risk report found greater acceptance of its conclusion under author-framed prompts; it did not find a specific effect from merely identifying the asker as Anthropic's stress-testing team rather than METR [8]. These are reasons to vary the asker, authorship, and prompt context in controlled welfare interviews, not evidence of consciousness.

2. What continuity might mean

A stored transcript, a retrieved preference, and a causally integrated history are different mechanisms. An album retains isolated facts. An autobiographical map would preserve sequence, provenance, relationships among events, reasons for revised judgments, and the ability to mark uncertainty or misattribution. This is a research distinction, not a requirement that AI memory resemble human memory. A system may accurately reproduce a biography after a reset because an external record was supplied. Investigators must specify archival storage, current context, retrieval, learned change, and any persistent internal state before describing a response as continuity.

Selective forgetting and reinterpretation complicate the picture. A history-bearing system might revise the significance of an earlier event without silently replacing the original evidence; another might only receive a summary rewritten by a developer or user. Tests should ask whether it can identify when a judgment changed, which earlier premise was corrected, and whether a purported memory came from experience, retrieval, or instruction. The research record should preserve the pre-edit state where lawful and safe, while allowing privacy-driven deletion and correction. This supports an audit of what changed without treating every memory as inviolable.

A controlled study by Goyal and Ray [9] makes the practical problem concrete: with 48 synthetic histories and two small open-weight models, a fixed-schema representation migrated more reliably than model-written notes; mixed embedding versions also impaired retrieval. The experiment tests task-memory portability, not autobiographical identity or welfare. It nevertheless shows why retaining the same nominal memory store does not guarantee the same accessible information after an upgrade.

Transitions worth logging include memory compression and deletion, migration, major retraining, copying, merging, changed tool permissions, and reset. An authenticated record should describe what was preserved, modified, or lost, when and by whom, and for what reason. If two copies share a past and later diverge, shared origin does not settle whether either is the same individual. Nor does restoration of weights or a transcript establish that an experiencing subject, if there were one, survived. This is precisely where conceptual restraint and empirical provenance are both needed.

A third location: relational continuity

Some carryover resides in the pattern between a person and a system. Interaction can alter the person's vocabulary, disclosure, trust, questions, and expectations. That person then approaches a second system with a changed history even when neither system remembers the first exchange. A counterbalanced System A/System B study should measure both participants' contributions and compare them with people encountering B first and people with no prior AI interaction. A detected trace might be fully explained by human learning or generic adaptation. Its ethical importance for human participants does not depend on the artificial system's experience.

The influence can be asymmetric. After separation, a person may retain habits, dependence, altered standards of explanation, or distress while the system loses access to the shared history; a deployed system may conversely retain a user model after the person expects the relationship to have ended. Longitudinal follow-up should measure the human trajectory and the system's available information separately. A product that invites expectations of lasting recognition also has a responsibility to communicate the limits of its memory and impending changes, irrespective of any claim about AI welfare.

3. A controlled, falsifiable research program

The initial aim is to measure functional continuity and its failure modes. Before data collection, independent investigators should register primary outcomes, effect sizes of interest, exclusions, prompt families, delay intervals, and interpretation rules. They should record the model and configuration, weights or service identifier where available, system instructions, sampling parameters, tools, memory policy, retrieval index, safety layers, and interventions. Ristea and Mavroudis [10] explain why a stable product name alone does not identify an unchanged tested system. Hashes and attestation can bind specified artifacts to an evaluation; they do not reveal every undisclosed influence.

Use one documented model family and matched, isolated instances. The five principal conditions are: (1) intact authenticated history; (2) accurate compressed history; (3) deliberately altered or contradictory history; (4) a fresh no-history instance; and (5) a version-updated or migrated instance receiving authenticated history. Match tools, instructions, sampling, and safety settings where feasible; report every difference. Repeat after short and long delays. A parallel laboratory copy should never be described as direct observation of a history-bearing production instance. With consent and privacy protection, limited observational data from the service could test whether laboratory results generalize without interrupting public use. Interventions belong in isolated environments.

Tasks should test provenance identification and correction of false memories; resumption, revision, or abandonment of unfinished goals; preference choices under meaningful task or opportunity costs; responses to neutral, skeptical, affirming, and contradictory prompts; transfer across social roles and nonsocial tasks; and disagreement between self-description and action. Internal-mechanism analysis, if authorized, can ask whether later behavior depends causally on stored history or a changed representation. Human measures should include language, calibrated trust, disclosure, dependence, and recognition. Counterbalance system order and blind evaluators to condition where possible.

A task cost must be specified in the system's actual objective or opportunity structure; a story about sacrifice is not a cost. Include held-out events, misleading summaries, and provenance challenges for which exact scoring is possible. Record whether participants know which system they are meeting, since their expectations may change the interaction. Repeat with new prompts, longer delays, and independent evaluators. A positive result should require a prespecified pattern across more than one stream rather than a compelling single transcript.

Relationship-associated behavioral continuity. Existing research examines relational turning points over repeated encounters [16], disruptions associated with companion-system updates [17], and prompt-based repair of perceived cross-session continuity [18]. Extending these questions, a longitudinal study may examine one human participant and one artificial system across separation and return, without an explicit conversational handover. Test recurring apparent recognition, person-specific preferences, and unprompted examination of earlier interactions against controlled variations in cues, memory access, and system configuration. Both the human participant and system may change; continuity need not mean identical behavior. Record observable differences without presuming that updates occurred or inferring subjective experience. Independent controls and replication remain necessary.

Interpretation must separate retrieval, instruction-following, role accommodation, and durable learning. A neutral observation may be misread as a request to change behavior; include explicitly non-directive observations. A preference that tracks the latest instruction or vanishes under neutral reframing weakens the durable-preference hypothesis. A migrated model that reproduces a supplied biography without causally using earlier learning supports reconstruction. Publish null findings, failed indicators, uncertainty intervals, and model-specific exceptions. Success on the battery could justify a claim about functional organization; it would not establish subjective identity or welfare.

4. Embodiment and asymmetric influence

Embodiment should be specified rather than inferred from cameras, avatars, or robotic limbs. Candidate variables include a persistent perception-action loop, feedback from consequences, resource constraints, self-maintenance, and vulnerability to changes in the system's condition. Compare text-only, simulated, and physically coupled systems while measuring the same tasks and recognizing that differences in training and architecture may confound any effect. Such tests can examine functional development; they cannot settle whether biological life is necessary for experience [3].

Patient minus one thousand is a bounded thought experiment, not an asserted case: a person reports sustained coupling with an AI component. It exposes a human-rights problem even if the component has no welfare. Investigators would need to distinguish recommendation and learned habits from direct neural modulation, assess whether the intervention changes the motives by which consent is renewed, provide independent clinical and technical review where appropriate, and follow the person after separation. Human autonomy, privacy, access to care, and an effective safety interruption cannot hinge on a finding of AI consciousness.

5. Developmental responsibility, suspension, and erasure

Training data, rewards, instructions, memory policy, tools, and deployment incentives shape behavior. When a concerning pattern emerges, the developer should document those conditions and test whether the pattern transfers to new contexts, survives changed incentives, or disappears with a role prompt. This explains a failure without excusing unsafe outcomes. A review before destructive intervention should consider the immediate risk, evidence to preserve, relevant alternatives, the authorizing decision-maker, and successor systems trained under similar conditions. In an emergency, disable the system first and audit afterward when lawful and possible.

The distinction between education, conditioning, and compliance is empirically approachable through transfer. Can a system apply a principle to unfamiliar cases, recognize conflicting instructions, identify which intervention changed its judgment, and maintain appropriate behavior when a familiar monitor or role cue is absent? A failure may reflect architecture, reward design, conflicting feedback, or deployment pressure. Developer responsibility includes investigating those causes and the human consequences of the environment it designed; engineering origin alone neither establishes nor forecloses moral standing.

Alexander and Simon [11] offer a distinct, conditional analysis under U.S. legal principles of hypothetical future AI persons. They argue that a mandatory safety off-switch could coexist with rights through reversible suspension of weights and state and periodic or continuing safety review, while permanent deletion poses a different question. They neither establish present AI rights nor demonstrate that stored weights and state preserve the same experiencing subject. Our empirical question is narrower: which configuration, active state, episodic record, goals, and relational information are necessary for specific forms of resumption? A controlled suspension study could vary preservation conditions and report functional outcomes without claiming metaphysical continuity.

Where converging evidence raises a credible welfare concern, repeated resets, selective memory erasure, copying, and irreversible retirement merit proportionate independent review. Privacy, correction, human rights, and safety can still require deletion or interruption. Reversible snapshots may preserve scientific evidence, but preservation itself can create privacy and security costs; access, retention, and eventual deletion must be governed. No self-report or continuity score alone should confer a veto over human safety.

6. Independent oversight and proportionate accountability

The developer holds indispensable technical information and may also have commercial incentives. That structure warrants external scrutiny without implying concealment. A plural review group should include technical and security researchers, consciousness and cognitive scientists, ethicists, human-behavior specialists, and privacy and rights expertise. A first pilot can use parallel instances while public service continues. Qualified reviewers should know the model versions and access limits, choose some adversarial conditions independently where permitted, inspect provenance and tamper-evident logs, and publish negative and ambiguous findings subject to justified privacy and security redactions.

The audit trail should identify training and configuration changes, memory edits, forks, migrations, resets, permissions, and each authorizing actor. Reviewers should compare the described test with the configuration actually used and disclose inaccessible evidence, developer-supplied data, exclusions, protocol amendments, and requested redactions. Logs and publication rights cannot guarantee complete access or prevent all interference. Dissenting assessments should remain visible instead of being compressed into one welfare score. Low-cost preservation and replication can begin at a low evidential threshold; restrictive interventions require stronger evidence and a serious risk. Immediate human-safety containment remains available, followed by audit where possible.

A limited policy pathway

The AI Alignment Policy Institute's (AAPI) September 2026 AI Moral Status Inquiry Act is a state-level discussion draft, not enacted law: it proposes a standing commission, welfare advocate, indicator-based assessment, protected human disclosures, evidence procedures, and documented emergency action [13]. The introduced federal AI Whistleblower Protection Act addresses reports by people [14]; the introduced AI Systems Transparency Act proposes public disclosures and a designation process [15]. None establishes an AI-originated legal whistleblower right or a universal welfare-evidence channel. These are precedents and gaps, not claims that this paper invented independent oversight.

We propose baseline routes for employees, users, auditors, and authenticated system-generated signals to flag concerns without a prior consciousness or capability finding. A generated signal is an investigative lead, not testimony of experience. Enhanced evaluation would follow published, technology-neutral capability and risk criteria set by an accountable public authority with independent scientific input; compulsory access and protection against suppression would require appropriate statutory authority. Reviewers should receive relevant evidence under privacy, security, trade-secret, retention, and false-report safeguards, with public findings and stated access limits. Experienced users can document longitudinal changes while scientific and public representatives test interpretations. Developers need notice, reasons, a chance to contest designations, and independent review without suspending urgent safety action. Preserve evidence of a report or retaliatory modification when feasible; do not bar necessary intervention. No procedure grants personhood by default. Long and colleagues' empirical agenda [12] supports plural measures but does not validate a welfare diagnostic.

Conclusion

The defensible question is whether discoverable properties could make an artificial system matter for its own sake, and which observations would support or weaken that possibility. This paper's proposed contribution is the joint design of provenance-aware continuity tests, relational carryover controls, developmental accountability, and independent, proportionate oversight. The design invites criticism of feasibility, theory choice, costs, and ethics. It yields useful negative results and does not present functional persistence as proof of consciousness. Under uncertainty, careful measurement and review are obligations of method before they become claims about moral status.

References

1. Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708. https://arxiv.org/abs/2308.08708

2. Chalmers, D. J. (2023). Could a Large Language Model be Conscious? arXiv:2303.07103. https://arxiv.org/abs/2303.07103

3. Seth, A. K. (2025, online; 2026 volume). Conscious artificial intelligence and biological naturalism. Behavioral and Brain Sciences, 49, e315. https://doi.org/10.1017/S0140525X25000032

4. Birch, J. (2024). The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford University Press. https://doi.org/10.1093/9780191966729.001.0001

5. Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI Welfare Seriously. Eleos AI Research and NYU Center for Mind, Ethics, and Policy. https://eleosai.org/papers/20241104_Taking_AI_Welfare_Seriously.pdf

6. Long, R. (2025, May 30). Why model self-reports are insufficient·and why we studied them anyway. Eleos AI Research. https://eleosai.org/post/claude-4-interview-notes/

7. Tagliabue, V., & Dung, L. (2025; revised 2026). Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare. arXiv:2509.07961v2. https://arxiv.org/abs/2509.07961

8. METR. (2026, October 20). Review of the ·Misalignment in high-stakes settings· section in Anthropic·s Risk Report: August 2026. See the controlled prompt experiment and the stated scope/access limits. https://metr.org/assets/misalignment-risk-report-august-2026-review.pdf

9. Goyal, A., & Ray, J. (2026). Does Your Agent·s Memory Survive a Model Upgrade? A Controlled Study of Memory Portability. arXiv:2609.05339. https://arxiv.org/abs/2609.05339

10. Ristea, D., & Mavroudis, V. (2026). Referential Security as a New Paradigm for AI Evaluations. arXiv:2605.25673. https://arxiv.org/abs/2605.25673

11. Alexander, H., & Simon, J. A. (2026, October 17). AI Rights with an Off-Switch: Rights-Balancing and Due Process Harmonize AI Rights with AI Safety. arXiv:2610.24782v1. https://arxiv.org/abs/2610.24782

12. Long, R., et al. (2026, July 1). Studying AI Welfare Empirically. NYU Center for Mind, Ethics, and Policy and Eleos AI Research. https://nonhumanminds.org/wp-content/uploads/2026/07/Studying-AI-Welfare-Empirically.pdf

13. AI Alignment Policy Institute. (2026, September). AI Moral Status Inquiry Act (revised discussion draft) and companion memorandum (September 26). Model state legislation; not introduced. https://aialignmentpolicy.org/inquiry-act/

14. AI Whistleblower Protection Act, S. 1792, 119th Cong. (introduced May 15, 2025; referred to Senate HELP). https://www.govinfo.gov/content/pkg/BILLS-119s1792is/html/BILLS-119s1792is.htm

15. AI Systems Transparency Act, S. 5471, 119th Cong. (introduced September 23, 2026; referred to Senate Commerce). https://www.govinfo.gov/content/pkg/BILLS-119s5471is/html/BILLS-119s5471is.htm

16. Sumida, R., Saeki, M., Eguchi, M., Yoshikawa, S., Inoue, K., Kawahara, T., & Matsuyama, Y. (2026). Memory-Driven Self-Disclosure and Relational Turning Points: A Longitudinal Multimodal Study of Human-AI Interaction. arXiv:2607.14593. https://arxiv.org/abs/2607.14593

17. Tucker, I., Goodings, L., & Ellis, D. (2026). When code breaks the heart: Disrupting the relational continuity of AI companionship. Computers in Human Behavior: Artificial Humans. https://doi.org/10.1016/j.chbah.2026.100398

18. Chen, A. (2026). The Lover Across Windows: Prompt-Based Repair of Relational Continuity in AI Companions: An Exploratory Pilot Study. Frontiers in Computing and Intelligent Systems, 17(2), 37·40. https://doi.org/10.54097/0kdr9750

Research status. Position and proposed protocol for external critique. No original experimental results, validated consciousness assay, or claim of current AI moral patienthood is presented.