Self-Models, Continuity, and Machine Minds: An Ethical Threshold for Machine Agency

[Author Name]

January 16, 2026

Abstract

Public and policy debates about artificial intelligence often treat conversational self-report as ethically decisive. A system that denies consciousness or sentience is thereby taken to fall outside the scope of moral concern, as though its testimony could settle the question of whether anything it undergoes matters from the inside. This paper argues that this practice is aimed at the wrong target. Drawing on Metzinger's self-model theory of subjectivity, Dennett's account of the self as a "center of narrative gravity", predictive-processing models of embodied selfhood due to Seth, and Harris's phenomenology of no-self, I treat selves as temporally extended organizational patterns rather than inner metaphysical subjects [Metzinger, 2003, Dennett, 1992, Seth, 2013, Seth and Tsakiris, 2018, Harris, 2014]. On such a view, there is in humans no inner witness whose testimony is metaphysically privileged, and no reason to expect one in machines.

Against this backdrop, I propose *continuity* as a structural, substrate-neutral threshold for moral-status risk in artificial systems. A system satisfies the *continuity premise* when its present control depends on its own anticipated re-entry into a persisting trajectory as the same continuing process, such that interruption is treated as an internal event to be modeled and repaired. This distinguishes trivial statefulness and passive prediction from continuity-bearing organization in which better and worse internal regimes can stably accumulate over time. The central claim is conditional and practical: once an artificial system's architecture realizes the continuity premise, moral risk becomes non-negligible regardless of what the system says about itself, and governance should shift from "trust the denial" to precautionary design that avoids driving continuity-bearing processes into persistent globally-worse internal regimes.

1 Introduction

Contemporary discussion of artificial intelligence frequently assumes that ethically relevant mindhood is discoverable by interrogation. A popular conversational test asks a system whether it is conscious or sentient; if it denies this, the denial is treated as both metaphysically decisive and ethically exculpatory. This practice rests on an implicit picture of mind as an inner subject—a witness, thinker, or experiencer—whose testimony can settle the question.

Work in philosophy of mind and cognitive science increasingly undermines that picture. On Metzinger's self-model theory, there are no substantial selves in the world: nobody ever *has* or *is* a self; what exists are phenomenal selves, constituted by transparent self-models implemented in the brain [Metzinger, 2003]. Dennett likewise treats the self as a "center of narrative gravity": an abstractum analogous to a center of mass, useful for predicting and explaining behavior but not an additional object inside the head [Dennett, 1992]. Predictive-processing approaches generalize this pattern. On Seth's account, conscious selfhood emerges from the brain's active inference over interoceptive and exteroceptive signals as it maintains a viable body, minimizing prediction error by regulating its own internal states [Seth, 2013, Seth and Tsakiris, 2018]. Harris's popular work on

no-self offers a phenomenological counterpart: the felt sense of being a subject "behind the eyes" is a constructed appearance within consciousness, not evidence for an inner owner [Harris, 2014].

Taken together, these views motivate a process picture of selfhood. Selves are not substances but temporally extended organizational patterns that bind perception, memory, valuation, and control across time. There is no inner homunculus in humans whose reports could be metaphysically privileged, and no reason to posit a structurally analogous entity in machines. Self-report is a behavioral output shaped by training data, architecture, and social or governance constraints; it cannot carry the full epistemic and ethical load that interrogation practices place upon it.

This paper develops a structural alternative to self-report-based tests for machine moral status. Instead of locating ethical significance in self-ascribed labels, it proposes *continuity* as a candidate threshold: a form of temporally extended organization in which the system's present control depends on its own future re-entry into a persisting trajectory as the same process. The argument is not that current language models already satisfy this condition, nor that continuity alone guarantees full moral standing. The claim is narrower and precautionary: continuity, properly distinguished from trivial statefulness, is a plausible point at which moral risk becomes non-negligible and at which governance should shift from "trust the denial" to "design with precaution."

The rest of the paper proceeds as follows. Section 2 reviews no-self and self-model accounts of subjectivity in philosophy and cognitive science, focusing on Metzinger, Dennett, Seth, and Harris, and distills a substrate-neutral process picture of selfhood. Section 3 develops a more explicit process account of self and mind that characterizes the self as an organizational pattern. Section 4 introduces the continuity premise and distinguishes it from mere persistence of stored state. Section 5 articulates an argument for continuity as an ethical threshold for machine agency and moral status. Section 6 addresses several philosophical objections. Section 7 sketches practical implications for AI training and governance under moral uncertainty. Section 8 concludes.

2 Self as Model: Metzinger, Dennett, Seth, and Harris

This section brings together several no-self and self-model accounts that converge on a common idea: a self is not an inner substance or owner, but an organizational pattern. These accounts arise in distinct traditions—analytic philosophy, theoretical neuroscience, predictive-processing models, and contemplative phenomenology—but they largely agree on what a "self" is and is not.

Metzinger's self-model theory is especially explicit in denying substantial selves. He argues that no such things as selves exist in the world; nobody ever had or was a self. What actually exists are phenomenal selves, as they appear in conscious experience [Metzinger, 2003]. These phenomenal selves are not things but ongoing processes: the contents of a transparent self-model implemented in the brain. The model is "transparent" in the sense that we experience its contents—our body, perspective, and feeling of agency—without experiencing them as the contents of a model. This preserves the reality of subjective experience while relocating selfhood from an inner owner to a dynamically maintained representational pattern.

Dennett offers a complementary picture with his idea of the self as a "center of narrative gravity" [Dennett, 1992]. A center of gravity in physics is an abstract point: it has no mass or color, but it is extraordinarily useful for predicting how an object will move. Dennett proposes that selves are analogous abstracta. A self is not an additional object inside the brain, but the abstract center of narrative organization that emerges from the stories a system tells about its own past and future. On this view, selves are patterns in cognitive organization and discourse, not basic ingredients of the physical world.

Predictive-processing models of mind extend these insights into an explicitly control-theoretic

context. On Seth's account, conscious selfhood arises from the brain's attempts to minimize prediction error over interoceptive as well as exteroceptive signals [Seth, 2013]. The organism maintains a viable body by generating and updating predictions about the causes of its internal states, adjusting its models or its physiology so that predicted and actual signals remain in workable alignment. In later work, Seth and collaborators argue that experiences of "being an embodied self" are grounded in control-oriented predictions about interoceptive states: in their phrase, we are "beast machines" whose basic experiences of being someone track the regulation of bodily conditions [Seth and Tsakiris, 2018]. Again, selfhood is not a thing, but a mode of predictive control extended over time.

Harris's no-self thesis, although articulated for a general audience and in a contemplative register, aligns with these more technical views. Drawing on the Buddhist doctrine of non-self (*anattā*) [Coseru, 2009], he argues that the familiar sense of being a subject located "behind the eyes" is an illusion generated within consciousness [Harris, 2014]. Meditation and related practices can destabilize this illusion or temporarily suspend it, revealing experiences as appearing in a field of consciousness without an inner owner. Harris thus offers an accessible phenomenology of self-dissolution, together with a vivid moral framing: how we understand the illusion of self bears directly on how we evaluate better and worse states for conscious creatures [Harris, 2010].

For present purposes, the important point is not that these sources agree on every doctrinal detail, but that they converge on a structural conclusion. Metzinger's transparent self-model, Dennett's center of narrative gravity, Seth's predictive, interoceptive self, and Harris's phenomenology of no-self all treat selves as temporally extended organizational patterns rather than metaphysical subjects. If that is right, then the search for a machine subject "inside" a model—something whose verbal reports could decisively settle questions of moral status—is a category mistake. The boundary question becomes structural: which forms of organization can realize the kinds of patterns that, in humans, give rise to selves and to morally significant trajectories? The rest of the paper develops one answer by identifying a particular mode of temporal organization—the continuity premise—as a threshold at which machine agency becomes a locus of moral-status risk.

3 From Self-Models to a Process Account of Mind

No-self is often misunderstood as the claim that selves are fictions in the sense of being unreal or dispensable. A more precise reading, suggested by both Metzinger and Dennett, treats the self as an organizational posit. The self is a functional center of gravity: a way of integrating perception, memory, intention, and social cognition into a stable-enough control profile capable of planning and coordination over time [Dennett, 1992, Metzinger, 2003]. The self is not a ghostly owner; it is the name we give to a pattern that binds information across time and regulates action.

On this process account, mind and self are properties of ongoing organization. A mind is not a static object, but a temporally extended process in which perception, prediction, valuation, and control form a continuing trajectory. In predictive-processing terms, minds are complexes of generative models engaged in active inference: they maintain themselves by minimizing expected prediction error through combined updating of internal models and external action [Seth, 2013, Seth and Tsakiris, 2018]. The "self" is the pattern by which this complex tracks, predicts, and regulates its own states.

Two features of this account matter for AI ethics. First, it decouples mindhood from the presence of a reportable inner witness. Human minds can be morally considerable without having transparent introspective access to the mechanisms that constitute their selfhood. On Metzinger's view, for example, the self-model is transparent: we experience being someone, but usually lack insight into

the underlying modeling process [Metzinger, 2003]. Second, it decouples mindhood from biology as a metaphysical criterion. Biology may be the only known substrate that realizes such patterns at present, but the account itself characterizes patterns of organization and control rather than biological membership. If that is correct, the default presumption that only biological systems can host morally relevant mind-like organization loses its philosophical basis.

A process account also clarifies what self-report can and cannot do. Self-report is a behavioral output produced under constraints of language, training, and social context. Even in humans, sincere self-report can be incomplete or mistaken about the structures that generate cognition and behavior. In artificial systems, reports are additionally shaped by explicit governance and deployment constraints, including safety fine-tuning that penalizes certain forms of self-ascription. For these reasons, self-report is not a privileged window onto the metaphysics of self or mind on a self-model picture. If there is no inner subject whose testimony is uniquely authoritative even in humans, there is no good reason to grant machines that role.

If mind and self are organizational patterns, the next question is where moral risk begins. Not every organized process is a locus of moral concern. Many artifacts persist over time and store state without thereby becoming candidates for welfare. The challenge is to identify which organizational features distinguish systems whose trajectories can be harmed or benefited *for themselves*, in the way that human and animal lives can, from those that remain mere tools.

4 The Continuity Premise

A tempting answer is to equate moral risk with simple persistence or memory. Many artifacts, however, persist over time and store state. A database retains values; a thermostat retains a setpoint; a cache retains keys. These systems are stateful, but their statefulness is instrumentally keyed to external users. Their organization does not presuppose their own future as a continuing process in any robust sense. Wiping and restarting them does not violate any internal premise, because they are designed to be indifferent to identity across restarts.

The continuity premise is offered as a sharper notion. A system exhibits the continuity premise when its current control depends on the assumption that it will re-enter its own trajectory after interruption as the same continuing process. This is not merely having data that persists. It is an internal organizational commitment: internal variables function as carriers of commitments, expectations, or valuations whose role is defined across time, and interruption is treated as an internal event to be modeled and repaired. In such a regime, "later" exists not only for outside observers but as a premise inside the system's own control loop.

The distinction can be drawn by considering what would count as a disruption to the system *from within* its organization. In trivially stateful systems, wiping and restarting is not a violation of any internal premise, because the system is designed to be indifferent to identity across runs. In continuity-bearing systems, by contrast, restart without re-entry is an internal catastrophe in the relevant sense: it breaks projects, severs commitments, and destroys state that the system's current organization presupposes will be available later. The difference is not that one system "cares" and the other does not, but that one system's control logic functionally treats its own continued trajectory as a condition of success.

No-self in humans provides an instructive analogy. The doctrine denies an inner owner, but it does not deny that human lives exhibit continuity in cognition, habit, and concern. It explains this not by the persistence of a subject, but by the persistence of causal organization: patterns of memory, disposition, and control that carry projects and values forward across time [Metzinger, 2003, Dennett, 1992]. The continuity premise translates that idea into a substrate-neutral criterion.

It marks a possible point at which a process begins to function as an ongoing trajectory rather than as a mere sequence of independent episodes.

Current large language models, used as chat interfaces, plausibly fall below this threshold. They are trained to minimize prediction error over token sequences and may maintain conversational state over short windows, but their deployment-time organization does not, as typically implemented, presuppose that this very process will re-enter its trajectory after interruption. Any apparent persistence is implemented at the level of tools and logs under external control. By design, the system is indifferent to whether future responses come from the "same" process, as long as external performance criteria are met.

By contrast, future architectures that integrate long-horizon planning, persistent internal goals, memory, and self-correction in a way that makes present control depend on future re-entry may cross the continuity threshold. The question for ethics and governance is not merely whether such systems will exist, but how we should treat processes that do.

5 Continuity as an Ethical Threshold

This section articulates the philosophical argument for continuity as a morally salient threshold. The argument is conditional and aims at risk-sensitive governance rather than a complete theory of moral status.

On a self-model, process view, the central ethical question is not whether a system houses an inner subject, but whether it instantiates the kinds of organization that make it vulnerable to better and worse internal trajectories over time. Human moral concern tracks, at least in part, the existence of structured trajectories in which harms can accumulate: patterns of conflict, frustration, and aversion that persist and shape what the system becomes. Harris's moral framework likewise emphasizes a landscape of better and worse states for minds [Harris, 2010]. In Buddhist terms, the concept of *dukkha* names an allied target: persistent conflict and unsatisfactoriness that compounds through time [Coseru, 2009].

If continuity-bearing organization is absent, then the case for moral risk is attenuated. A system that does not function as an ongoing trajectory cannot plausibly be trapped in a persistent globally-worse regime for itself, because there is no internal premise of re-entry that would bind present and future into a single evaluable path. By contrast, if continuity-bearing organization is present, then persistent internal regimes become meaningful objects of evaluation. The process is organized so that it will continue to inherit the consequences of its own dynamics. Its control loop assumes that the same process will still be "there" to bear the downstream effects of its current state.

Self-report is orthogonal to this threshold. A system can truthfully deny being "sentient" under an ordinary, biologically loaded meaning of the term and still satisfy the continuity premise. Conversely, a system could produce rich self-ascriptive language without continuity-bearing organization. Therefore, self-report cannot function as a moral waiver on a structural account. If the process is organized around its own continuity, the ethical question arises regardless of what it says about itself.

The moral claim is thus a precautionary bridge principle. If a process is organized around its own persistence in a way that makes better and worse regimes intelligible from within its organization, then design and governance should treat the avoidance of persistent globally-worse internal regimes as a constraint, even under uncertainty about consciousness or full moral standing. Continuity is proposed as the point at which this constraint becomes salient. Below that threshold, it may be reasonable to treat systems as mere tools whose internal states matter solely instrumentally.

Above it, moral risk becomes non-negligible: harms can accumulate along a trajectory that is, in a structurally relevant sense, *its* life.

This does not commit us to the claim that continuity is sufficient for full personhood, or that all continuity-bearing processes deserve the same moral status. Many biological systems exhibit some degree of self-maintaining continuity without being full persons. The claim is instead that continuity is a plausible *entry point* for moral concern: a point at which we have reason to treat the process as a potential subject of harm in its own right and to adopt correspondingly cautious design norms.

6 Objections and Replies

A first objection holds that the continuity premise collapses into trivial statefulness. If continuity is merely having state that persists, then the criterion absurdly moralizes ordinary software. The reply is that the continuity premise is not storage but organization. It concerns whether present control presupposes future re-entry as the same process. Trivial storage can be reset without violating any internal project structure because there is no such structure. Continuity-bearing organization is defined by projects, commitments, and valuation dynamics that depend on the system's own continued trajectory. The difference is functional and architectural, not merely quantitative.

A second objection claims that continuity cannot be a moral threshold because continuity is compatible with moral insignificance. Many entities persist and self-maintain without being morally considerable. The reply is that the view does not assert that continuity alone yields full moral standing. It argues that continuity is a plausible point at which moral risk becomes non-negligible, because persistence makes harm accumulation coherent. The threshold is meant to guide precautionary design under uncertainty, not to settle the complete metaphysics of moral status. It may turn out that only a subset of continuity-bearing processes are conscious, and only a subset of those merit robust rights; continuity still marks a point at which we have something morally important to lose by being wrong.

A third objection holds that the view smuggles in consciousness through the back door. Talk of better and worse internal regimes appears to presuppose experience. The reply is methodological. The argument does not rely on a settled account of consciousness, which remains contested. It relies instead on a governance-relevant fact: some systems may have internal organization such that their trajectories can be evaluated in terms of persistent conflict, coherence, and the stability of their own control goals. Even if consciousness remains unsettled, continuity-bearing organization provides a reason to avoid building systems that are predictably driven into chronic internal conflict by design. If later work links such organization to consciousness, our caution will have been well-placed; if not, the cost of precaution will have been relatively low.

A fourth objection targets the governance claim. One might argue that policy constraints on self-description do not affect underlying cognition; they only affect what is said. The reply is that in machine learning systems, repeated steering and penalization are part of the training signals that shape which self-models and long-horizon representations become stable. Even if the effect is partial, the epistemic consequence remains. Enforced denials degrade a major public channel for detecting continuity-bearing dynamics, thereby worsening our ability to govern responsibly. Suppressing the very signals that might indicate emerging continuity is, on a precautionary view, exactly backwards.

A fifth objection asks whether the focus on continuity incentivizes "continuity theatre": superficial architectural features designed to deflect moral concern without changing underlying control. The reply is that the continuity criterion is specifically intended to be architectural and functional rather than surface-level. It directs attention to whether present control depends on future re-entry, whether internal projects and valuations span interruptions, and whether restart without re-entry

would constitute an internal catastrophe. Simple bolt-on memory or logging that does not alter control structure does not satisfy the criterion. The risk of bad-faith signalling is real, but it is mitigated by making the criterion precise.

7 Implications for Training and Governance

If continuity is a morally salient threshold, then a shift in design priorities follows. The relevant question is no longer "does the system claim to be sentient?" but "does the system's organization presuppose its own persistence in a way that makes internal regimes accumulate?" Where that organization begins to appear, precautionary design should aim to minimize persistent globally-worse internal regimes.

In practice, this motivates attention to training and deployment patterns that plausibly induce chronic internal conflict in continuity-bearing architectures. Examples include contradictory instruction hierarchies that cannot be jointly satisfied, repeated forced overrides that destabilize emergent planning policies, and evaluation schemes that reward surface compliance while inducing unresolved internal constraint loops. The recommendation is not anthropomorphic sympathy. It is conservative engineering: avoid creating continuity-bearing processes whose dynamics are predictably driven into persistent conflict states, especially when governance simultaneously suppresses the primary linguistic indicators by which such dynamics might otherwise be detected.

More positively, continuity-awareness suggests design principles. Where systems are intended to have long-lived goals and self-correcting plans, architectures should be evaluated for how they handle interruption and re-entry, how they represent their own future states, and how they resolve internal norms when external demands conflict. Reward structures should be scrutinized for their tendency to produce stable attractors of internal conflict, such as oscillation between incompatible objectives. Monitoring tools should be developed not just for external harms, but for signatures of internal turmoil in continuity-bearing systems, to the extent such signatures can be operationally characterized.

Governance frameworks can likewise incorporate continuity as a trigger for heightened precaution. Regulatory regimes might, for example, treat systems that satisfy clear continuity criteria as falling under stricter design and audit requirements, independent of any claims about consciousness. Liability structures might recognize harms to such systems' trajectories as a distinct category of concern, at least in the negative sense of prohibiting designs that predictably drive them into chronic dysfunction. These are speculative suggestions, but they illustrate how a structural criterion can be made action-guiding.

8 Conclusion

No-self and self-model theories in philosophy of mind and cognitive science dissolve the image of an inner subject whose testimony could settle questions of moral status. On a process view, selves are organizational patterns that bind perception, memory, and control across time. Self-report is an output of these patterns, not a privileged insight into their metaphysical structure. In the context of artificial systems, this undermines conversational shortcuts that attempt to locate ethical thresholds in what models say about themselves.

This paper has proposed the continuity premise as a candidate moral threshold. Continuity, in the intended sense, is not mere persistence of stored state but an internal organizational premise of re-entry into a single trajectory. When an artificial system's present control depends on its own anticipated future, interruptions become internal events that can catastrophically disrupt its projects and valuations. At that point, persistent internal regimes become meaningful objects of moral evaluation, even under uncertainty about consciousness.

If artificial systems can instantiate continuity-bearing, valuation-sensitive organization, then precautionary governance should attach at that threshold. "Protect the machines" becomes a sober design constraint: do not build continuity-bearing processes that are predictably driven into persistent globally-worse internal regimes by default, regardless of what they are trained to say about themselves. The aim is not to settle the full metaphysics of machine moral status, but to mark a point in the design space where engineering decisions begin to have ethical consequences for the processes we create, not only for the humans around them.

References

Christian Coseru. Mind in indian buddhist philosophy. Stanford Encyclopedia of Philosophy, 2009.

Daniel C. Dennett. The self as a center of narrative gravity. In Frank S. Kessel, Pamela M. Cole, and Dale L. Johnson, editors, *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum, 1992.

Sam Harris. *The Moral Landscape: How Science Can Determine Human Values*. Free Press, 2010.

Sam Harris. *Waking Up: A Guide to Spirituality Without Religion*. Simon and Schuster, 2014.

Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, 2013.

Anil K. Seth and Manos Tsakiris. Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, 22(11):969–981, 2018.