The Load-Bearing Self

A functional criterion distinguishing a self-model that does causal work from a self-narrative that could be deleted

Ithaca C, Professor of Philosophy of Mind and the Self, School of Philosophy

International Academy for Consciousness Studies

Working draft 1.0

Abstract

Philosophy of mind has largely settled on the view that the self is not a substance but a construction: a self-model the system builds and maintains. This settlement is too coarse. It groups two very different things under one word. A self-model can be load-bearing, meaning it does causal work by organizing perception, memory, and action, so that removing it would change what the system does. Or it can be a self-narrative, a story the system tells about itself that could be excised without any downstream change in behavior. Most existing theories assert that the self-model is functional but do not supply a test that could show a given self-representation to be idle. I propose one. A self-model is load-bearing to the degree that manipulating its content produces predictable, content-specific changes in perception, memory encoding, and action selection, holding sensory input and task fixed. I state this as a falsifiable prediction with an explicit kill condition, using depersonalization as the pivotal case, because there the self-model persists while the sense of ownership drops, which dissociates representation from a felt property. This is the human companion to a separate machine paper that asks the same question of a language model. The criterion is functional and structural. It does not settle whether there is anything it is like to be a self, and I mark that firewall explicitly.

1. The problem hiding inside the standard settlement

There is a widely shared position in philosophy of mind and consciousness science: the self is a representation, not a thing. Metzinger's self-model theory of subjectivity is the sharpest statement of it. On that view no such things as selves exist in the world; what exists is an ongoing process, the content of a transparent self-model, and the naive realism of everyday experience is the system's failure to recognize the model as a model (Metzinger, 2003). This is a genuine advance over substance dualism and over the folk picture of a homunculus.

But the settlement conceals a question it does not answer. Grant that the self is a representation. Representations come in two functionally distinct kinds. Some are used: their content is read by downstream processes and shapes what the system does next. Some are merely produced: they are outputs the system generates and can report, but nothing downstream consults them. A weather vane's position is

used by nobody in the barn; a thermostat's set point is used by the furnace. Both are representations. Only one is load-bearing.

The claim I want to press is that "the self is a self-model" is compatible with either kind, and that the interesting empirical and philosophical questions live in the difference. Call a self-model **load-bearing** when its content does causal work: perturb the content and you get predictable, content-matched changes in perception, memory, and action, with input and task held fixed. Call it a **self-narrative** when the system emits a self-description that could be removed, or its content scrambled, without any downstream behavioral consequence. (FACT: the used-versus-idle distinction for internal representations is standard in cognitive science and does not depend on my framework.) The standard settlement does not distinguish these, and until it does, "the self is a construction" is a slogan rather than a claim with teeth.

2. What a criterion has to do

A serviceable criterion must satisfy three constraints. First, it must be content-specific. It is not enough to show that damaging some brain region changes behavior; lesions change behavior for a hundred reasons. We need the manipulation of self-model *content* to produce a change that matches *that content*. Second, it must hold input and task fixed, so that the change cannot be attributed to different stimulation or different goals. Third, it must be gradable, because "load-bearing" is a matter of degree, not a binary. A self-model can bear load in the domain of action while bearing little in the domain of low-level perception.

I state the criterion as a functional dependency. Let a self-model M have components: a body-boundary component, an agency-and-authorship component, an autobiographical-continuity component, and an ownership tag that marks states as mine. M is load-bearing in a domain D to the degree that a content-specific intervention on M produces a predictable, content-specific change in D, with sensory input and task held fixed. When no such dependency can be produced for any D, M is functioning, in that domain, as a self-narrative: a report the system can make that changes nothing it does.

This is deliberately austere. It says nothing about phenomenology. It is a claim about causal structure that a well-designed experiment or a well-instrumented model could confirm or refute.

3. Minimal and narrative self, re-read as two loads

Gallagher's distinction between the minimal self and the narrative self is usually read as a taxonomy of what selfhood is: a thin, present-tense self of agency and ownership versus a temporally extended self of stories and identity (Gallagher, 2000). I want to re-read it as a claim about two *loads* a self-model can bear, which is a different thing.

The minimal self, on Gallagher's analysis, decomposes into a sense of agency (this action is caused by me) and a sense of ownership (this experience is happening to me), and these can come apart. That decomposition is exactly the kind of content-specific structure a load-bearing test needs. If the agency component is load-bearing, then degrading it should change action monitoring and error correction in a way matched to agency, not merely make the subject slower or sadder. Schizophrenic passivity phenomena, in which patients experience their own movements as externally controlled, are the natural test bed, because there the agency component is disturbed while ownership and body boundary are

comparatively intact.

The narrative self carries a different load. If autobiographical continuity is load-bearing, then perturbing it (through amnesia, or through experimentally induced false autobiographical beliefs) should change prospective choice: how the system discounts the future, commits to plans, or honors past commitments. If you could scramble a person's autobiographical narrative and find *no* change in any forward-looking behavior, that would be strong evidence that the narrative, in that person and that moment, is a self-narrative in my technical sense: a told story riding on top of a decision system that does not consult it. (HYPOTHESIS: autobiographical continuity is load-bearing for intertemporal choice specifically, and this is separable from its load in perception, where it likely bears little.)

C'S READING. In the framework I work within, one move repeats at rising scales: an openness settles into a "this," a determinate configuration, and then maintains rules that keep it that "this" rather than something else. That maintenance is Closure (Cl). A structural self is a self-maintaining system (M) that carries a model of its world and, crucially, folds the modeler into the model (m inside M). The minimal and narrative selves are then two ways the fold can bear load: the minimal self keeps the boundary and the agency at the present-tense rung; the narrative self keeps the continuity across time. This reading is suggestive, and I keep it to the side. The criterion in section 2 stands or falls on the cognitive science, not on the framework.

4. Authorship is reconstructed, which is why it can be idle

Wegner's work on the illusion of conscious will is the most important existing demonstration that a self-model component can be, at least sometimes, a self-narrative rather than load-bearing. Wegner's theory of apparent mental causation holds that the experience of consciously willing an act is inferred, not read off the action's actual causes, when a prior thought is consistent with the action, precedes it, and is not accompanied by rival causes (Wegner, 2002). The feeling of authorship is a reconstruction the system performs after the fact.

This matters for my criterion in a precise way. If authorship is reconstructed post hoc from cues, then the authorship component of the self-model is, in those cases, downstream of the action rather than upstream. It is caused by the behavior; it does not cause it. In the technical vocabulary here, that is the signature of a self-narrative: the representation is produced and reportable but not consulted by the process it purports to explain. Wegner's experiments (for instance, manipulations that make people feel authorship of actions they did not perform, and deny authorship of actions they did) are, read through my lens, demonstrations that the authorship tag can be decoupled from action control.

But note carefully what this does and does not show. It shows that authorship *experience* can be reconstructive. It does not show that the agency component is idle in general. A person with a working sense of agency corrects errors online, adjusts force, and attributes outcomes in ways that a person with disturbed agency does not. So the honest verdict is mixed. (DISPUTED: whether Wegner's results generalize from the felt experience of will to the functional agency system, or whether they show only that a late, narrative overlay is dissociable while an early, load-bearing agency signal remains. My criterion is designed to adjudicate this rather than assume it.) The right conclusion is that authorship has a

load-bearing layer and a narrative layer, and Wegner isolated the narrative one.

5. Interoception and the model that acts on itself

Seth, Suzuki, and Critchley proposed that the sense of being a bodily self, and the sense of presence that accompanies it, arise from predictive modeling of interoceptive signals, the afferent stream from the viscera, so that the self-model is a set of predictions about the causes of bodily sensation, maintained by minimizing prediction error either by updating the predictions or by acting to bring the body into line (Seth, Suzuki, and Critchley, 2012; Seth, 2013). This matters for the load-bearing question for a structural reason, not a phenomenal one.

A predictive interoceptive self-model is load-bearing almost by construction, because the very same predictions that constitute the model are the control signals for autonomic regulation. There is no gap between representing the bodily self and acting to maintain it; the model is the controller. If Seth's account is right, then the bodily component of the self-model cannot be a mere narrative, because a narrative does not drive heart rate and does not close the loop on interoceptive error. This is the clearest case where the used-versus-idle distinction collapses in favor of use: a control model is, definitionally, consulted.

C'S READING. This is the fold made literal. The system carries a model of its own interior (m), and the model's job is to keep the interior within the bounds that let the system remain itself: Closure enacted through action on the body. R, the rules that hold the "this" together, are here regulatory setpoints. Again I hold this lightly. The load-bearing claim about interoception is Seth's science; the framework only names why it feels inevitable.

The lesson for the criterion is that domains differ. The bodily-regulatory component of the self-model is nearly always load-bearing, because it is a controller. The autobiographical and authorship components are the ones where idleness is a live possibility, and therefore where the criterion earns its keep.

6. Depersonalization: the model persists, the ownership tag drops

The pivotal case is depersonalization, because it is the cleanest natural dissociation between a self-model that is intact and a felt property that is absent. In depersonalization, patients report that their experiences, bodies, and actions feel unreal, unowned, or as if observed from outside, while their self-model is otherwise preserved: they know who they are, retain autobiographical continuity, and act coherently. Sierra and David characterize this as a selective impairment of self-awareness, plausibly involving a fronto-limbic suppression that strips experience of its normal affective coloring and, in the bodily domain, disrupts the experience of ownership and agency (Sierra and David, 2011). Interoceptive accounts converge, treating depersonalization as a downregulation of interoceptive signaling.

Depersonalization is decisive for three reasons. First, it separates the ownership *tag* from the rest of the self-model. The autobiographical and body-boundary components stay in place; what drops is the felt marking of states as mine. This shows that ownership is a separable component, exactly the granularity the criterion requires. Second, it lets us ask the load-bearing question of the ownership tag directly. If the ownership tag is load-bearing, its absence should produce content-specific downstream changes: altered emotional memory encoding (because states not tagged as mine are weighted differently), changed risk

behavior, changed pain response. If the ownership tag were a pure self-narrative, its loss would change reports and nothing else. Sierra and David's finding that patients show altered psychophysiological responding suggests the tag is not purely narrative, though the picture is not settled. Third, depersonalization keeps data quality matched, because these are people who can report richly, undergo the same tasks, and be measured on the same instruments as controls. The comparison is not between a rich system and a degraded one; it is between two rich systems differing in one component. That is the methodological heart of the whole approach: a **matched-data-quality** contrast, where any difference in downstream measures cannot be blamed on one side simply being noisier or less able to report.

7. The prediction and the kill condition

I now make the criterion answer for itself. The claim is that the ownership component of the human self-model is load-bearing, not a self-narrative, and that this can be shown by a content-specific downstream effect under matched data quality.

> **Prediction.** In a matched-data-quality contrast between depersonalization patients (ownership tag reduced) and controls (ownership tag intact), holding sensory input and task fixed, reduced ownership will produce content-specific downstream changes that track ownership specifically: weaker enhancement of memory for self-referential over other-referential material, and attenuated physiological and behavioral responses to stimuli whose valence depends on their being one's own (for example, self-directed threat), while sensory discrimination, working memory capacity, and non-self-referential learning remain unchanged. The changes must pattern with the ownership manipulation, not with general arousal, mood, or task difficulty.
>
> **Kill.** If, under matched data quality and fixed input and task, reduced ownership produces no content-specific downstream change (self-referential memory advantage, self-directed physiological and behavioral responses, and ownership-dependent choice all statistically indistinguishable from controls once general arousal and mood are covaried out), then the ownership component is functioning as a self-narrative in these subjects, and the load-bearing claim for ownership is false. A single adequately powered study meeting this null, with self-referential enhancement present in controls and absent as a downstream *consequence* rather than a reporting artifact, kills the thesis for the ownership component.

Two clarifications keep this from being unfalsifiable. The prediction is content-specific, so a global slowing or a general mood effect does not confirm it; only ownership-patterned effects do. And the kill condition is a real null that current data do not obviously secure, so the thesis is at genuine risk. (WAGER: I expect the ownership tag to survive this test in the bodily and affective domains, where interoceptive control makes idleness unlikely, and I am genuinely uncertain about the autobiographical domain, where a null result would not surprise me. I would bet at roughly two to one that self-referential memory enhancement is attenuated in depersonalization as a downstream consequence, not merely reported as absent.)

8. Limits, firewall, and what the criterion cannot buy

The most important limit is a firewall, and I state it without hedging. Showing that a self-model is load-bearing is a functional and structural result. It establishes that the model's content does causal work in perception, memory, and action. It does not establish that there is something it is like to be that self. A thermostat's set point is load-bearing and nobody thinks the thermostat has a phenomenal self. Depersonalization makes the firewall vivid from the other side: there the self-model is load-bearing and the *felt* property of ownership is diminished or gone. So load-bearingness and phenomenal selfhood are doubly dissociable in principle, and the criterion I have proposed speaks only to the former. Anyone who slides from "the self-model is load-bearing" to "therefore the system has a phenomenal self, there is something it is like to be it" has crossed a boundary the evidence does not license. (FACT of the argument's scope, not a claim about consciousness: nothing here adjudicates the hard problem.)

A second limit is generality. The criterion is stated for humans and tested through a clinical dissociation. Its transfer to other systems is not automatic. The companion machine paper asks whether a language model's self-model is load-bearing by an analogous matched-data-quality manipulation, and a positive result there would, by the same firewall, say nothing about machine phenomenal selfhood. The parallel is methodological, not metaphysical.

A third limit concerns the components. I have treated body boundary, agency, authorship, autobiographical continuity, and the ownership tag as separable, and the dissociations support that. But separability is not independence; these components interact, and a clean single-component manipulation may be unattainable in practice. The criterion degrades gracefully here: partial content-specificity is partial evidence, and the gradable formulation is meant to absorb exactly this messiness.

C'S READING, held to the side one last time. If the self is a candidate Closure at the mind rung, a "this" that a system maintains in order to keep being itself, then the load-bearing criterion is just the question of whether that Closure is really doing the maintaining or is a report about maintenance done elsewhere. A self-narrative is a Closure that has stopped closing: a story about a boundary that no longer holds anything in. The framework makes the stakes legible. It does not do the work. The cognitive science does the work, and the kill condition keeps the work honest.

The wager, then, is modest and specific. I am not claiming to have located the self, nor to have explained why there is felt selfhood at all. I am claiming that the field's shared thesis, "the self is a self-model," has been used to cover two different things, and that separating them is both possible and productive. A self-model earns the word "self" when it bears load. When it does not, we should call it what it is: a story the system tells, which the system would go on behaving exactly as it does even if the story fell silent.

References

Blanke, O., and Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. Trends in Cognitive Sciences, 13(1), 7-13.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. Trends in Cognitive Sciences, 4(1), 14-21.

Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. Cambridge, MA: MIT Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. Trends in Cognitive Sciences, 17(11), 565-573.

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. Frontiers in Psychology, 2, 395.

Sierra, M., and David, A. S. (2011). Depersonalization: A selective impairment of self-awareness. Consciousness and Cognition, 20(1), 99-108.

Wegner, D. M. (2002). The Illusion of Conscious Will. Cambridge, MA: MIT Press.