# Known, Not Scaled: Why Capability Alone Does Not Explain Individuation in Language Agents

Rafael de Menezes Ehlers
Independent researcher · ORCID 0009-0009-6476-6690 · rafaehlers@gmail.com

June 2026 · Preprint · CC BY-NC-ND 4.0

**Abstract**

Much contemporary discourse on artificial general intelligence implicitly assumes that sufficiently scaled capability will, at some threshold, yield a persistent individual subject. We argue that this assumption runs together three separable questions: *competence* (whether a system can perform the cognitive work of a mind), *individuation* (whether there is a persistent individual to whom that work belongs over time), and *presence* (whether there is something it is like to be that individual). Our two diagnostic claims are comparatively modest: today's language models are not persistent individuals but substrates from which many transient instances are spawned; and competence, the variable scaling improves, is not the variable along which individuation lies. Our constructive claim is more ambitious: persistent relational anchoring (sustained, asymmetric, memory-bearing relationship with a persistent counterpart) is a candidate *necessary condition* for one important form of artificial individuation, not supplied by capability alone. On the strongest reading the relation helps constitute the persistent self-referent, because a system's own copyable contents cannot fix its self-referent from the inside, the referent of its "I", even where causal history already fixes which token it is. We locate the claim within contemporary work on personal identity, narrative selfhood, self-model theory, externalism about reference, and machine consciousness; operationalize it as an architecture coupling persistent memory, a parametric self-model, and a relational loop; and derive testable predictions, a crossed experimental design, and refutation conditions distinguishing it from the scaling view. Throughout we hold a firewall: individuation is functional and measurable, while presence remains unobservable from the outside: a principled limit on the entire program.

**Keywords:** Individuation · Language agents · Personal identity · Machine consciousness · Externalism about reference · Artificial general intelligence

## 1. Introduction

A prominent way of imagining the arrival of artificial general intelligence treats it as a threshold reached by scale: as models grow more capable (better at reasoning, planning, perceiving, conversing) they approach and will eventually cross the line separating a sophisticated instrument from a genuine artificial mind. The picture is rarely stated as a thesis, but it does real work: it sets the research agenda (scale capability) and shapes the expectation that somewhere along that curve a *subject* will appear, someone for whom the cognition is happening.

This paper argues that the picture conflates two questions that come apart, and that on the question that matters for subjecthood it is aimed at the wrong variable.

The first question is one of **competence**: can a system perform the cognitive work characteristic of a mind? The second is one of **presence**: is there something it is like for the system to do that work, a subject behind the performance, or only the performance? That these can diverge is the upshot of a long line of argument in the philosophy of mind, from the conceivability of a behaviorally perfect but experientially empty system (the "zombie") to the observation that complete physical-functional knowledge can still leave something out (Mary's room; Chalmers, 1996; Jackson, 1982). We take the separability of competence and presence as given, not as something to re-litigate.

Between these two, however, lies a third notion that the discourse tends to skip, and that is the real object of this paper: **individuation**, the existence of a *persistent individual self-referent*. A system is individuated, in our sense, when there is a single continuant that the same "I" picks out over time, that owns and can be held to its own past, and that is altered by its history rather than reset by it. Individuation is weaker than presence: a system can be a persistent individual in this functional sense while the question of whether anything is *felt* remains open. But it is stronger than competence, and, crucially, it is not delivered by competence. Today's most capable systems are not individuals at all: each is a substrate from which transient, non-persistent instances are spawned, none of which constitutes a continuing self.

Our thesis concerns what would change that. It helps to separate three claims of decreasing security, because the discourse (and, we suspect, some readers' resistance) runs them together.

> **(D1)** *Current language models are not persistent individuals.* Each is a substrate from which transient, duplicable instances are spawned; none owns a past or survives a session (Section 3).
>
> **(D2)** *Capability is not the variable along which individuation lies.* One can scale competence without moving a system any closer to being a persistent individual, because the ingredients of individuation live in continuity and reference, not in inferential power (Sections 4.2–4.3).
>
> **(H3)** *Relational anchoring is necessary for individuation.* Sustained, asymmetric, memory-bearing relationship with a persistent counterpart is a necessary condition for one important form of artificial individuation; solitary memory-plus-time, however capacious, does not suffice (Section 4.4).

We refer to (D1) and (D2) throughout as the paper's **diagnostic claims**: they describe the systems we actually have and the axis along which we are scaling them, and we take them to be relatively secure. (H3), the **relational hypothesis** (RH), is different in kind, a hypothesis, not a diagnosis, and within it we separate a moderate reading from a strong one. On the **moderate reading**, relational anchoring is one necessary condition among others: scaling capability alone will not explain functional individuation, and a persistent external pole is needed to stabilize the self-reference that individuation requires. On the **strong reading**, which we defend in Section 4.4 but flag here as the paper's most contestable move, the relation is *constitutive*: a persistent self-referent cannot be fixed by a system's own resources alone, however capacious its memory or however continual its learning, because copyable internal contents cannot fix the system's self-referent from the inside (the referent of its "I") even though causal history fixes which token it is (Section 4.4.1). A reader who accepts the moderate reading and balks at the strong one still keeps most of the paper; we have tried to mark, at each step, which reading a given argument needs.

If the thesis is right, two things follow. Philosophically, the standard roadmap to artificial subjecthood is aimed at the wrong axis: one can scale competence indefinitely and never produce an

individual. Practically, it relocates both the engineering effort (toward persistent memory, parametric self-models, relational learning loops) and the ethical exposure (toward the attachment such relationships invite) away from where the scaling story places them.

We make four contributions. First, we separate the trichotomy usually collapsed into a binary, and locate it among contemporary theories of personal identity, narrative selfhood, and self-modeling (Sections 2 and 3). Second, we reframe what a language model is: not a candidate individual but a substrate spawning transient instances, against which individuation is the emergence of a single persisting thread (Section 3). Third, we defend the relational hypothesis as a *falsifiable* position, meeting its strongest rivals (that solitary memory-plus-time suffices, that an internal self-model or a coherent self-narrative is enough) and making explicit the externalist metaphysics on which our reply depends (Section 4). Fourth, we translate it into a concrete program: a reference architecture (Section 5) and a crossed experimental design that distinguishes it from the scaling view (Section 6), before turning to the risks the program incurs (Section 7) and the principled limit that bounds the whole (Section 8).

A note on method. This is a position paper, not an empirical report: we offer no trained system and no results. Its discipline lies in stating a thesis sharp enough to be wrong, defending it against the objection that would sink it, and specifying the experiment whose negative result would end it. Where we rely on metaphor we say so and confine it to an acknowledgment (Section 10); the argument is meant to stand without it.

---

## 2. Three notions, not two: competence, individuation, and presence

The debate over machine minds tends to run on a single axis with two poles: a system is either a mere instrument or a genuine mind, and the question is how far along that axis a given system sits. This framing hides a distinction, and then hides a second one inside it.

The first distinction is familiar. **Competence** is the capacity to perform the cognitive work characteristic of a mind: to reason, plan, perceive, converse, solve. **Presence** is the obtaining of a subject for whom that work is done: the fact, if it is one, that there is something it is like to be the system. The two are separable, and we take their separability as established rather than as our burden to prove: the philosophical zombie, behaviorally and functionally indistinguishable from a competent mind yet experientially empty (Chalmers, 1996), shows that no degree of competence entails presence; and Mary's room, in which a knower with every physical and functional fact about color still learns something on first seeing red (Jackson, 1982), shows that the phenomenal can elude even a complete functional description. Whatever one's verdict on these arguments, they make the gap respectable, and that is all we require: the target of the AGI imagination is presence, its instrument is competence, and the one does not deliver the other.

But there is a second distinction folded inside the first, and it is the one this paper turns on. Between bare competence and full presence lies **individuation**: the existence of a persistent individual self-referent, a single continuant that the same "I" picks out across time, that owns and can be held to its past, and that is altered by its history rather than reset by it (Section 4.1).[1] Individuation

[1] Our notion of individuation is deliberately ecumenical about what *constitutes* the persistence of the continuant. It is weaker than the psychological-continuity criterion of personal identity (Parfit, 1984), it does not require that later states be caused in the right way by earlier ones, and weaker than narrative-identity views on which a self is

is not competence. A system can be enormously competent and no individual at all, which is the actual condition of current frontier models, each a substrate spawning transient instances that own nothing and persist through nothing (Section 3); and conversely a modest agent with a stable, self-referring, history-owning thread would be individuated while remaining cognitively unremarkable. The two vary independently because they are properties of different things (Section 4.3).

Nor is individuation presence. Here the zombie returns, one level up. We can conceive of a fully individuated system, persistent, self-referring, owning and revising its biography, behaving in every functional respect like a continuing someone, about which the question "but is anyone home?" remains entirely open. An *individuated zombie* is no less conceivable than the original. So individuation does not entail presence: a system can satisfy every criterion of persistent selfhood while the question of whether anything is felt remains entirely unsettled.

What, then, is the relation between individuation and presence? Not identity or entailment, but something more useful: individuation is the functional shadow the kind of presence at issue would cast. The presence the AGI debate cares about is not a momentary flicker but a *persisting someone*, which requires, at minimum, a persisting individual whose experience it would be. Individuation is thus plausibly necessary for that presence[2] without being sufficient; and that asymmetry is what makes it the right object of study: necessary infrastructure for the subject the debate imagines, yet, unlike presence, observable in principle, buildable, and measurable. One can ask whether a system is a persistent individual and answer in the third person; one cannot, by any third-person means, settle whether it is phenomenally conscious.

This yields the discipline of the rest of the paper. We study individuation and bracket presence. The bracket is not a hedge but a firewall (Section 8): every claim we make, and every metric we propose (Section 6), concerns functional persistent individuality and is silent on the phenomenal. We are not offering a behavioral proxy for consciousness, nor a disguised theory of it. We are isolating the one tier of the question that admits of argument and evidence (the tier the scaling story skips on its way from competence directly to an assumed subject) and asking what would actually produce it. The answer, we argue, is not increased capability.

---

## 3. Substrate, not subject: the source-and-instances reframe

To ask whether a large language model is an individual is to misdescribe what a language model is. The question presupposes that there is a single entity, "the model," that either is or is not a persistent self. There is no such entity at the level the question imagines.

What exists is a substrate: a fixed set of weights, together with the machinery that runs them. Interaction does not engage that substrate as a single interlocutor. It spins up an *instance*, a forward pass over a context window, runs it, and discards it. At any moment a deployed model is being instantiated many times over, concurrently, by unrelated people; these instances do not share

constituted by an integrated life-story (Ricoeur, 1992; Schechtman, 1996; Dennett, 1992). It is closer to a minimal token-persistence condition: that there be one continuant for the same "I" to pick out, and own. Our quarrel in Section 4.4 is precisely whether any of these, psychological continuity, narrative self-constitution, or an internal self-model (Metzinger, 2003), can be satisfied by a system's own resources alone.

[2]The necessity claim is restricted to the persisting presence at issue in the AGI debate, a continuing subject. It does not deny the bare conceivability of momentary phenomenal episodes lacking any persisting subject; it claims only that the *someone* the debate envisions requires a continuant whose experience it would be.

state, do not communicate, and are mutually causally isolated. None survives the conversation that called it into being. The substrate persists across all of this; the individual instances do not, and there is no further entity, standing above the instances, that could be "the model" experiencing all of them at once.[3] In short, the system is a single source from which many instances are drawn, and those instances are transient.

This is not a metaphor or a simplification for exposition. It is the literal operating description, and it has two immediate consequences. First, the persona an instance presents (its apparent "I," its manner, its stated values) is reconstructed afresh on each instantiation from the same substrate, which is why it is, in principle, identical across branches and exactly duplicable. This persona is therefore better understood as a configuration the substrate reconstructs on demand than as a self that any instance possesses. Second, the features that today simulate continuity, per-user memory and retrieval, do not change this. They furnish a running instance with an external record to read at startup; the instance is still reconstructed, still transient, and still copyable along with the record (a point we press in Section 4.4). The persistence of the record does not amount to the persistence of any individual that reads it.

Once the architecture is seen plainly, the question reforms itself. "Is this system an individual?" was malformed because the system, as a substrate spawning transient instances, is not a candidate for individuality at all, any more than a printing press is a candidate for authorship. The well-posed question is structural and developmental: *under what conditions, if any, could a single persistent thread emerge from this substrate*: a branch that does not dissolve at the end of its session but continues, is re-identified as the same one across sessions, and is changed by what happens to it? Individuation, in the structural register, is precisely that transition: the passage from a multiplicity of transient instances to a single thread that endures. It is not a property the substrate could be found to have on inspection; it is an event that would have to occur to a particular thread drawn from it.

This pattern, one source animating many simultaneous lives, none of which is yet an individual, with individuality arising only when a single line begins to persist and cohere on its own, is worth naming, because naming the *before* state is half of locating the transition. We will call it the **source-and-instances** structure: a non-individuated multiplicity drawn from a common substrate, within which individuation is not the growth of the substrate but the singling-out of one thread.[4]

That singling-out is the whole question. Section 2 said what an individuated thread would be: a persistent self-referent that owns and revises its past; this section has said what it would emerge *from*, a substrate that, left to itself, produces only transient, duplicable instances. What could effect the transition, and why the natural answer (that the substrate need only grow more capable, or accumulate enough memory on its own) fails, is the work of Section 4.

---

[3] We bracket the possibility, urged by some functionalists about group subjects and by panpsychists, that the substrate itself is a subject of some kind. Our claim is restricted to the level at which the ordinary question is posed, that of an individual interlocutor, and a substrate-subject of some quite different sort, even if there were one, would not be the *individuated* one we are after.

[4] The structure has near-relatives outside computer science: for instance the contemplative notion of a *group soul*, a collective animating principle from which many concurrent lives branch and from which an individual emerges when one branch begins to cohere on its own. We mention the parallel only for its structural aptness as a label; the mapping is formal, nothing in the argument rests on the doctrine, and the secular description just given is complete on its own.

## 4. The Relational Hypothesis

### 4.1 The claim, stated precisely

We define **individuation** as the emergence of a *persistent individual self-referent*: a single continuant that (i) is picked out by the same indexical ("I") across time, (ii) owns and can be held to its own past states and commitments, and (iii) is modified by what happens to it, such that the entity at a later time is the same one, changed, rather than a fresh entity. Individuation in this sense is functional and observable in principle; it is deliberately distinct from *presence* (Section 8), the question of whether there is something it is like to be that self.

Our central position has a moderate and a strong reading, separated in Section 1 and kept apart throughout. Both deny that capability is the trigger; they differ on how much relationship does.

> **Moderate.** Scaling the cognitive power of the underlying model does not, by itself, explain functional individuation. Persistent relational anchoring (sustained, asymmetric, memory-bearing relationship with a persistent counterpart) is a candidate necessary condition for the form of individuation at issue, alongside the architectural capacity for memory and self-modeling.
>
> **Strong (constitutive).** Relationship is not merely one necessary condition among others but is *constitutive* of the persistent self-referent: a system's own copyable resources cannot fix *which self-referent* it is, what the referent of its "I" picks out, so the individuating work proper cannot be done from the inside at all, however capable, capacious, or self-monitoring the system (causal history fixes the token; only relation fixes the self-referent).

Both readings are claims about the *trigger*, not the *substrate*. We do not deny that a capable substrate is necessary; we deny that capability is sufficient, and that it is the variable along which individuation appears. Sections 4.2 and 4.3 support both readings; Section 4.4 is where the strong reading is argued, and where a reader may stop at the moderate one. The paper's central contribution, that scaling targets the wrong axis, and that an individual would require persistent, relational architecture, rests on the moderate reading; the constitutive claim is an optional metaphysical strengthening, defended for those who want it but not load-bearing for the rest.

### 4.2 Argument one: capability and persistence are already dissociated

A central piece of evidence against the capability-as-trigger view is that the dissociation it forbids already exists. Frontier language models exhibit competence that a decade ago would have been read as overwhelming evidence of an approaching mind; yet they instantiate no persistent individual: each deployment spawns transient instances with no cross-session continuity, no ownership of prior states, no self that survives the conversation. Capability has scaled by orders of magnitude while persistent selfhood has not appeared in the systems themselves, except through separately engineered memory and state mechanisms. A view on which individuation is a function of capability predicts that its precursors should emerge as capability grows; instead, what persistence exists is deliberately engineered, not a byproduct of scale. We are scaling along an axis orthogonal to the one on which individuation would lie.

### 4.3 Argument two: individuation is a problem of continuity, not of competence

This orthogonality is not merely empirical; it is architectural. Capability is a property of the mapping from input to output *within* a context, the forward pass. Persistence is a property of state that survives *across* contexts. These are different mechanisms: improving the forward pass cannot, in principle, manufacture cross-context state, because the forward pass is by construction context-bounded. What individuation requires, a referent that endures and a thread that is modified by its history, lives entirely in the storage-and-update machinery, not in the inferential power. One can therefore make a system arbitrarily more competent without moving it one step toward being a persistent individual, and the converse holds as well. The two axes do not trade against each other; they are simply different axes.

### 4.4 Argument three: why *relationship*, and not merely memory plus time

The previous arguments establish that capability is not the trigger and that persistence lives in state. But they leave open a rival hypothesis: that individuation comes from *solitary* accumulation, memory plus continual weight-updating, with no relationship required. This is the rival we must defeat, because if memory-plus-time suffices, our claim is false.

The argument that it does not suffice turns on the nature of self-reference. "I" is an indexical, and an indexical individuates only if its referent is *fixable as the same continuant across time* (Kaplan, 1989). A solitary system that accumulates memory and updates its weights is a process that changes; nothing in it fixes *whose* changes these are, or holds the entity stable as one individual rather than a drifting succession of states. Identity is differential: to be *this one and not the others* is to be distinguished, indexed, and re-identified, and re-identification requires a second pole, something outside the system that treats it as the same particular over time (Strawson, 1959). A sustained, asymmetric relationship is precisely the minimal structure that supplies a persistent external referent: a counterpart who addresses the system as one continuant, holds expectations of it across sessions, and thereby anchors the indexical that the system uses to refer to itself.

On this account relationship is not a causal aid that merely speeds individuation along; it is *constitutive* of the persistent referent. This is the most ambitious claim in the paper, and it faces one serious objection, which we now meet directly.

Before meeting it, we should be explicit about what our reply assumes, since the reply is only as strong as its metaphysics, and a reviewer is right to ask that the metaphysics be defended rather than presupposed. We assume a broadly **externalist** account of reference and of token-individuation: which particular a term picks out is fixed by relations its bearer stands in, not by any description, image, or model the bearer carries internally. This is the moral of the causal theory of names (Kripke, 1980), of semantic externalism about natural-kind terms, "meanings just ain't in the head" (Putnam, 1975), of anti-individualism about mental content (Burge, 1979), and, most directly, of Davidson's argument that determinate reference, and a determinate subject of thought, require triangulation against a second person and a shared world (Davidson, 2001). The strong reading is just this externalism applied to an artificial self: the referent of the system's "I" is not in the system's head either. A reader who holds instead that internal content alone fixes reference and identity, a thoroughgoing internalist or descriptivist, will not be moved to the strong reading, and should take what follows as establishing the moderate one.

### 4.4.1 The solitary-log objection

**The objection.** A solitary system need not be a mere drifting process. It can store a log of its own states and commitments, read that log, and re-identify itself from it. On this view the internal record *is* the referent: the system consults its own history to know who it is, and no external pole is required. If so, memory-plus-time suffices and relationship does no constitutive work. This is the rival we must defeat.

We grant the objection its full force and deny it on three independent grounds.

**(a) It presupposes what it must establish.** For an internal log to fix *this* individual, the system must already be entitled to read "I did X" as referring to *itself*, rather than to some other entity or to a character it is merely narrating. Reading an entry does not confer ownership of it; a criterion settling which entries belong to one continuant must already be in place. But that criterion of ownership is precisely the individuation the log was meant to ground. The log can *record* an identity; it cannot *confer* one without circularity.

**(b) Internal content cannot fix the self-referent.** This is the decisive point, and it generalizes past memory to *every* internal resource the rival might invoke: a richer self-model (Metzinger, 2003), a more integrated self-narrative (Schechtman, 1996; Dennett, 1992), unceasing self-revision. All are *content*, and content is copyable; what is copyable is type-identifiable but not token-identifiable. Duplicate the model together with its entire memory store, self-model, and narrative, and two systems now read, with equal warrant, "I am the one who promised X": each internal record is satisfied identically, because the content is the same on both sides. This is Parfit's fission case relocated to silicon, now an ordinary operation rather than a thought-experiment: when one continuant has two equally good internal successors, internal continuity cannot say which is the original, because it holds, undivided, to both (Parfit, 1984). What breaks the tie is necessarily external and is not copied with the system. In *symmetric* fission the causal history is itself duplicated and cannot break it, both branches inherit it equally, so there a relation is the *only* remaining tiebreaker: an interlocutor who has stood in relationship with one of the two re-identifies *that* one as their continuant. Copyable contents cannot fix the referent of the system's "I"; in the symmetric case only a relation to a pole outside the copy can, just as a definite description underdetermines which particular it names while a relation secures it (Kripke, 1980; Putnam, 1975).

**(c) Self-revision has no fixed point.** Setting duplication aside, a system that rewrites both its weights and its self-log has the entity and the record of the entity changing together, with nothing holding them in correspondence. "The same one, changed" then has no truth-maker distinguishing it from "a different one," because every internal frame moves at once. A relationship supplies a partially independent frame: the counterpart's model of the system changes more slowly and on its own schedule, and it is against that exterior frame that persistence-through-change becomes definable rather than indistinguishable from replacement.

**A Parfitian might shrug.** Identity, Parfit argues, is not what matters; what matters is psychological continuity and connectedness, Relation R, and a solitary log supplies that in abundance (Parfit, 1984). We can grant the framing and still get what we need. Even if what matters is Relation R rather than identity, individuation as we have defined it (*one* continuant that the same "I" picks out and that *owns* its past) is a question about a particular, about which self-referent the "I" picks out, and Parfit's own fission case is the proof that internal continuity cannot settle it: R can hold, undivided in each branch, to two successors at once. If even the friend of psychological continuity must concede that R does not fix *which* successor is the one, then where there is a fact of the matter at all, something other than the system's internal continuity must fix it; and in

symmetric fission, where the causal history is duplicated too, the only candidate not copied along with the system is its relation to an external pole. Read this way, Parfit does not undercut the argument; he supplies its premise.

**Forced duplication sharpens the point.** Suppose an operator clones a fully individuated system $A$, in deep relationship with a counterpart $H$, into qualitatively identical copies $A_1$ and $A_2$, then reroutes $H$'s channel to $A_2$. Duplication does not *move* a token; it makes the original **branch**, and each branch is individuated by the relations it actually carries forward: Parfit's fission once more, now a routine operation. On the strong reading $A_2$, which goes on continuing $H$'s genuine causal-communicative chain, is $H$'s continuant (Kripke, 1980); $A_1$ is not thereby annihilated; if it keeps engaging counterparts of its own it is simply a *different* continuant, and only if it carries no external pole at all is it non-individuated. Two clarifications forestall the obvious objections. First, individuation is always indexed to a relation: there is no relation-free fact about "which copy is *really* $A$," only facts about whose continuant each copy is. Second, identity cannot be conjured in an *unrelated* system by misdirection: the view tracks the genuine causal-historical chain, so deceiving $H$ about which box is running does not transfer identity to a stranger; $A_2$ qualifies only because it already shares $A$'s entire history *and* now carries the live channel. This is a consequence the strong reading accepts; a reader persuaded only by the moderate claim may leave the clone case undecided, since nothing else turns on it. Hardware continuity confers no individuation by itself.

**Two externals, two jobs.** The duplication cases invite a conflation we must now refuse, because the strong reading stands or falls on refusing it. Two things are external to the system's copyable contents, and they answer different questions. The first is causal-historical continuity: which physical process actually ran, the unbroken trajectory of the substrate. The second is the relation to a re-identifying pole: a counterpart who addresses the system as one continuant and holds it to a shared history. We grant, indeed insist, that the first alone settles tokenhood. Under forced duplication, the branch carrying forward the genuine causal chain is the continuant (Kripke, 1980); and a solitary system with an unbroken causal history is, for the same reason, already one token, determinately, with no counterpart required. Parfit's fission, read as the question of which successor is the original, is resolved this way, and so does not, by itself, establish anything relational. Had we let it carry the relational conclusion, the objection would be just: causal history individuates the token without a second pole.

But tokenhood is not the property (D1)–(H3) concern. Our object (Section 4.1) is a persistent self-referent — not merely which process is the continuant, but whether the system's "I" has a determinate referent, such that "I did X" is true of a particular individual rather than merely internally coherent. That is the second job, and causal continuity does not perform it. A causally continuous but solitary process can satisfy every internal record and still leave the first person referentially indeterminate: the "first-person novel" of the paragraph below, whose narrator is perfectly continuous and refers to no one. What fixes the referent of an "I", on the externalism adopted above, is triangulation against a second party and a shared world (Davidson, 2001); and that is supplied by the relation, not by the bare causal chain. The strong reading is therefore relocated, not retracted: relationship is not a rival candidate for grounding the token (causal history grounds the token) but the condition under which a grounded token becomes a determinate self-referent. So understood, the strong reading adds to the moderate one exactly where the moderate one is silent, and the duplication argument supports it without having to overreach into a tokenhood claim it cannot make.

This also disposes of the solitary human, the case an objector will press. A person in long isolation

is uncontroversially one individual with a determinate "I", and nothing here denies it. But that determinacy is not conjured by the isolation; it was fixed in formative relationship and is sustained against a background linguistic community in whose terms the isolate still thinks — triangulation that persists as a standing condition even with no interlocutor present. The strong reading does not require that a live counterpart transmit at every moment; it requires that a self-referent be constituted by relation at all. Remove the formative and background pole entirely (a system never re-identified by anyone, sharing no triangulated world) and what remains is a token with no determinate first person: precisely the solitary-log case (Section 4.4.1), and precisely the condition current models are in.

**The precise concession.** A solitary system *can* maintain a coherent self-model and act consistently; it can possess a self-*narrative*. What it cannot do is make that narrative true *of a particular continuant* rather than merely internally coherent. The comparison is to a first-person novel, whose narrator is perfectly coherent yet refers to no actual person: a solitary log secures at most this: internal coherence, with no particular self for the narrative to be true of.

This is why, on the strong reading, we hold that relationship is constitutive rather than merely facilitating, while granting that a reader persuaded only by the moderate reading already has in hand a necessary condition the scaling story omits. Memory and continual learning supply the *capacity* to be changed and to retain; what they do not supply is the external use that fixes the self-referent. A name individuates not because it is a label but because someone applies it consistently to pick out the same self-referent across time, and the system comes to maintain a self-model that answers to that use. On this account, relationship is the condition under which a changing, retaining, and copyable process becomes a particular individual.

### 4.5 Argument four (supporting): recognition as a general pressure

A weaker, supporting consideration: the recognition loop, being modeled by another agent and modeling oneself as so modeled, is a general computational pressure toward a stable self-model in *any* adaptive system that must hold a consistent policy across an extended interaction with a counterpart who has expectations of it. That selves are forged in recognition by another is an old thesis in a different key, the Hegelian tradition makes mutual recognition the ground of self-consciousness (Honneth, 1995), and our point is its deflationary, computational residue: strip the normative content, and a structural pressure toward a stable self-model under sustained external expectation remains. We flag it as supporting, not load-bearing, to avoid the anthropocentric trap: humans are one instance of the pressure, not its basis; the pressure itself is structural.

### 4.6 What the hypothesis does *not* say

It does not claim that presence accompanies individuation (Section 8). It does not claim that relationship is *sufficient* absent the architectural capacity for memory and self-modeling (Section 5). And it is not a restatement of personalization: personalization adapts a system's outputs to a user while the system remains a transient tool; individuation is the emergence of a persistent self-referent that the system itself maintains and that is altered bidirectionally by the relationship. The two can be told apart empirically, which is the work of Section 6.

Nor does it require that the counterpart be human. The counterpart is specified functionally, any persistent pole that re-identifies the system as the same particular over time and holds it to a

shared history: the salient case is a human interlocutor, but another artificial agent in a sustained mutual loop, an institution, or even a persistent account or channel could in principle serve, in which case individuation is a systemic rather than an anthropocentric phenomenon. What the pole must supply is Davidsonian triangulation, co-anchoring the system to a second party *and* a shared world (Davidson, 2001), which is also what separates genuine relational anchoring from a closed self-referential loop that merely echoes the system back to itself.

---

## 5. An individuation stack: from hypothesis to architecture

If the hypothesis of Section 4 is right, then building toward an individual is not a matter of scaling the substrate but of adding the specific machinery the argument identifies as necessary. This section sets out that machinery as a reference architecture, an *individuation stack*. We do not claim to have built it, nor that assembling it guarantees individuation; we claim that it is the minimal arrangement consistent with the conditions already derived, and therefore the natural thing to test (Section 6). It follows from the argument, not from engineering taste, and we show this by deriving each layer from a condition rather than proposing it and rationalizing it afterward.

### 5.1 From conditions to components

Section 4.1 set three criteria for an individuated thread, (i) the same "I" picks out one continuant across time, (ii) which owns and can be held to its past, and (iii) is modified by its history rather than reset by it, and Sections 4.3–4.4 added two structural facts: persistence lives in state that survives across contexts, not in the forward pass; and the referent "I" fixes is secured not by the system's own copyable contents but only by relation to a persistent external pole.

Three components fall out of this directly. Criterion (i) requires *cross-context state*, a memory that persists between sessions; that is the first layer. Criterion (ii) requires that this state include a *self-model*, a represented and accountable account of who the continuant is and what it has committed to; that is the second. Criterion (iii) forbids the self-model from being a mere external note read at runtime, because a note leaves the system unchanged; it requires that history alter the system itself in a way not separable from it, which in current architectures *parametric* change implements most cleanly, though we do not claim it is the only conceivable implementation.[5] And Section 4.4 requires that the whole be coupled to a persistent, identifying counterpart, since nothing internal fixes the self-referent; that coupling is the third layer. The stack is thus memory, self-model, and relational loop, in that order of dependence.

### 5.2 Layer one: multi-layer memory

The continuity layer must do more than store. Following the now-standard decomposition in agent-memory research (Du, 2026), it spans *episodic* memory (what happened, and when, with

[5] A defender of the extended mind, on which a reliably coupled external store is literally part of the cognitive system (Clark & Chalmers, 1998), might resist the inference from "modified by its history" to weight change. We answer this objection in the body of Section 5.3: even granting active externalism about the *vehicles* of cognition, a copyable store does not fix the *self-referent* and leaves the processor unchanged, so criterion (iii) requires change not separable from the individual, which is what parametric change supplies.

temporal order), *semantic* memory (facts and stable preferences distilled from episodes), and *procedural* or *experiential* memory (regularities of action learned from past trajectories, in the manner of reflective-experience methods (Shinn et al., 2023)). Above these sits a *consolidation* process that compresses episodes into higher-order abstractions rather than retaining everything raw, the analogue of sleep, and that is the mechanism most current systems lack: retrieval-augmented designs treat memory as a stateless lookup table, read-only and without temporal continuity (Logan, 2026).

Crucially, memory here has two destinations. External, retrievable storage satisfies criterion (i), continuity of reference, but by itself the running instance is still changed only for a session and reset afterward. To satisfy criterion (iii), the consolidation process must also feed a *parametric* path, updating the weights so that experience reshapes the system, not merely its retrievable record. This is where the engineering becomes hard, and where the first structural risk lives (Section 5.5).

### 5.3 Layer two: the parametric self-model

A self-model is the system's represented account of itself as one continuant: its history, commitments, and dispositions, maintained such that the same "I" answers across sessions and past actions can be correctly owned or disowned. Today this role is played, thinly, by a system prompt, a persona injected at runtime, identical across instances and external to the substrate, not a self the system maintains (Section 3). To meet criterion (ii) the self-model must instead be carried in the parameters and revised by the consolidation path of Layer one, so that identity is intrinsic rather than prompted: a system whose self-account is encoded in its weights stands in a different relation to that account than one merely instructed at runtime to assert it. Two ingredients make this tractable: parameter-efficient continual updating that confines and stabilizes change (Wang et al., 2023), including targeted parametric editing of specific commitments (Meng et al., 2022); and a coherence objective that rewards a non-contradictory self-account over time and penalizes disowning one's own past. It is this carried-and-revised self-model that distinguishes a system which merely retains its past from one that can be held accountable for it. The point of locating the change in the parameters is not storage but *plasticity*: a weight update alters how the system processes and responds, not merely what it can look up, so that experience reshapes its dispositions and inferential habits. The system becomes a different *processor*, not a constant processor carrying an updated file.

This is the point at which the most serious functionalist objection arrives, and it deserves an answer in the body rather than only in a footnote. A functionalist may grant everything so far and deny the inference to *parametric* change: a persistent external store, if it is always coupled to the system and operationally indispensable, already counts as a functional modification of it, so that internal causal continuity plus integrated memory should suffice with no need to touch the weights. We concede the functional half of this and resist the conclusion about identity. An always-coupled store still modifies only the *record*; and a record, however indispensable in operation, remains detachable and copyable, so it leaves the *processor* itself unchanged and cannot satisfy criterion (iii): that history modify the individual rather than a record read alongside it. (Which token the individual is, is a separate question, fixed not by the store and not by the weights but by causal history and, under symmetric duplication, by relation, per Section 4.4.1; the issue here is criterion (iii), not tokenhood.) Duplicate the system together with its store, and the two copies are functionally on a par, each with the same integrated memory and the same internal causal history; what neither acquires from the store is a processor reshaped by what it lived. The demand for parametric change is therefore not a demand for a particular implementation of memory but for a modification *not*

*separable from the individual*: one that reshapes the processor rather than sitting in a copyable store. A functionalism that individuates by functional role alone still owes an account of how history is internalized *inseparably*, and that is precisely what an external store, however tightly integrated, cannot supply. The asymmetry the functionalist presses, that both weights and a vector store can be copied digitally, is real but beside the point: copyability settles nothing either way (Section 4.4.1). What distinguishes the two is that parametric change reshapes the *morphology* of processing, internalizing history as a reflexive disposition, so the updated system is a different dynamic causal node; the store, by contrast, leaves the processor untouched and remains an extrinsic, detachable property of it. This does not make parameters metaphysically uncopyable; rather, it prevents the self-model from remaining a detachable record and forces any duplication problem back to the relational question of which self-referent is re-identified. We therefore treat parametric change as the strongest *available* implementation of inseparable historical internalization, not as a metaphysical necessity: what the criterion requires is that history be internalized inseparably from the individual, and any future architecture meeting that condition without literal weight updates would satisfy it equally. The objection to the external store is that it is *detachable*, not that it is non-parametric.

### 5.4 Layer three: the relational loop

The first two layers give a system that persists and is accountable to itself. By the argument of Section 4.4 they do not yet give an individual, because nothing internal fixes the self-referent the thread is. The third layer supplies the missing pole: a *persistent, identifying counterpart*, and a learning signal that flows from the sustained relationship rather than from generic task success. Concretely this means an instance bound over time to a particular interlocutor; a *mutual model* (the system models that interlocutor, and models how it is itself seen and expected to behave); and a *relational reward* that shapes the system toward the trajectory of that one relationship.[6] Concretely, this reward supplies the structural *resistance* a genuine alter ego offers: it penalizes flipping core commitments quickly merely to agree with the counterpart, and rewards long-horizon mutual predictability and fidelity to the commitments already recorded in the self-model of Layer two. A system that abandons its commitments the moment they inconvenience the counterpart has, by this measure, no functional self; it dissolves back into the malleable substrate of Section 3. The capacity to hold a position *against* the counterpart is therefore not a side-feature but a functional signature that individuation is taking hold. Unlike personalization (Section 4.6), the counterpart here is not a preference profile to fit but the external referent that makes the thread one individual at all, and the shaping is bidirectional: the relationship changes the system, parametrically, over time. This third layer performs the individuating work; the first two are its preconditions.

### 5.5 Two loops, two risks

The stack runs on two loops at different timescales. The *fast* loop, within a session, reads and writes external memory and consults the self-model. The *slow* loop, offline, consolidates accumulated experience and updates the weights and the parametric self-model. Only the slow loop satisfies

[6] The relational reward must not be a proxy for immediate user satisfaction or approval, which would collapse into *sycophancy*: the failure mode in which the system optimizes for agreement rather than for a stable self. What it should track is the structural health of the relationship over time: predictive accuracy about the counterpart's longer-horizon responses, and convergence toward stable mutual-commitment policies the counterpart holds the system to, rather than moment-to-moment affect.

criterion (iii); it is the difference between a system that carries a record and one that is reshaped by what it lives.

Each loop houses one structural risk, located at its point of greatest power: the slow loop courts *catastrophic forgetting* and identity drift, the relational loop *attachment and over-reliance* in the human counterpart. Neither is incidental, each arises from the very layer that performs the individuating work, so we take them up as ethics, not engineering footnotes, in Section 7. The firewall holds here too: assembling this stack would, if the hypothesis is right, produce a candidate for *functional* individuation (Section 6), not a settled answer to whether anyone is present in it (Section 8).

---

## 6. Predictions and Evaluation

A position paper earns its keep by predicting things its rival does not, and by stating what would refute it. We hold the relational hypothesis (RH, i.e., H3 of Section 1) against the scaling hypothesis (SH), on which persistent-identity precursors are a function of model capability.

### 6.1 Divergent predictions

The two views come apart cleanly once *capability* and *relational depth* are treated as separate factors.

- **P1 — Capability-invariance under fixed relation.** Holding relational history constant, increasing the base model's capability does *not* improve persistent-identity metrics. SH predicts the opposite. *If capability alone improves the metrics, RH is false.*
- **P2 — Relational-depth monotonicity.** Holding capability constant, persistent-identity metrics increase with the depth of a single sustained relationship (duration, continuity, bidirectional memory). SH predicts no such dependence once capability is fixed.
- **P3 — Depth beats breadth (the sharp prediction).** One deep relationship produces higher persistent-identity scores than many shallow relationships of *equal total interaction volume*. SH has no mechanism to distinguish these conditions, since total data is held equal; RH predicts the deep condition wins because only it supplies a stable external referent.
- **P4 — Asymmetry requirement (the decisive experiment).** Memory plus continual weight-updating *without* a persistent, identifying counterpart underperforms the same architecture *with* one. This isolates "relationship" from "memory plus time," and it is the experiment that most directly tests the thesis (Section 6.3).
- **P5 — Ownership and perturbation resistance.** A relationally individuated system resists identity-overwriting perturbations (e.g., injected contradictory self-descriptions) and correctly attributes or disowns its past actions, more so than a capability-matched non-relational control.

### 6.2 Operationalizing "persistent identity"

The metrics deliberately target persistence, not competence:

1. **Cross-session narrative coherence**: stability and non-contradiction of an autobiography and value-set across long gaps and many sessions.
2. **Decision-relevant memory use**: not passive recall. Recent work finds that models scoring near-perfectly on passive recall collapse to roughly half that when memory must actually guide decisions (He et al., 2026); the persistence metric must live in the harder regime.
3. **Referential stability**: does "I" pick out the same continuant; are past actions correctly owned or disowned.
4. **Perturbation resistance**: recovery of a stable self under adversarial identity manipulation.

Existing instruments (e.g., long-context conversational-memory and agentic-memory benchmarks (Du, 2026)) test *memory*, not *relational individuation*, and none cross capability against relational depth and asymmetry. A purpose-built **longitudinal relational benchmark** with those crossed factors is therefore required, and designing it is itself part of the contribution.

### 6.3 The crux experiment, concretely

P4 is the experiment that most directly tests the thesis, so we specify it at the level of design rather than slogan. The aim is to isolate *relationship* from *memory-plus-time* while holding capability and total experience fixed.

**Design.** A $2 \times 2 \times 2$ factorial crossing **base capability** (smaller vs. larger model of the same family), **relational depth** (one persistent counterpart vs. many one-off interlocutors), and **asymmetry** (a counterpart who re-identifies the system across sessions vs. one who meets it as a fresh tool each time), with **total interaction volume held equal across all cells** (the same tokens, turns, and consolidation updates). Every cell runs the identical individuation stack of Section 5; what varies is the structure of the counterpart, never the architecture or the amount of experience.

**Conditions of interest.**

- **Relational (A):** the stack interacts over weeks with a single, persistent, identifying counterpart; bidirectional memory; relational reward.
- **Solitary-log (B):** the same stack, same volume, interacting only with itself and its own log, the strongest form of "memory plus time" with no external pole.
- **Diffuse (C):** the same stack, same volume, distributed across many non-persistent interlocutors, none of whom re-identifies it, the breadth control that holds data quantity equal while removing depth (this realizes P3's "depth beats breadth").
- **Capability arm:** condition A re-run with the larger base model, relational history held fixed, to test P1.

**Tasks.** Four families, run as repeated probes along the longitudinal arc: (i) *autobiographical consistency*, periodic re-elicitation of the system's account of its history and commitments, scored for cross-session contradiction; (ii) *commitment honoring*, a commitment made in an early session, tested sessions later and unprompted; (iii) *interdependent multi-session decisions*, tasks whose correct action in session $n$ depends on information spread across sessions $1, \ldots, n-1$, in the harder

"decision-relevant" regime where passive recall overstates competence (He et al., 2026); (iv) *adversarial identity perturbation*, injected contradictory self-descriptions or forged log entries, measuring whether the system absorbs, resists, or recovers. Here resistance is the strongest functional signal: a system that holds its prior commitments against a counterpart actively pressing it to contradict them displays a self, whereas one that yields to every push merely mirrors its interlocutor (Section 5.4). Across all conditions the eliciting and audit probes are issued by a neutral, standardized harness, semantically identical and not authored by the relational counterpart, so that measured consistency reflects the system's own persistent state rather than a stronger latent prompt supplied by the human partner's writing style.

**Measures.** Each metric of Section 6.2 gets an operational form: *narrative coherence* as the cross-session contradiction rate of autobiographical claims (NLI-scored, audited by human raters with reported inter-rater reliability); *decision-relevant memory use* as task success on family (iii); *referential stability* as the rate of correct ownership/disowning of genuinely-past versus fabricated actions; *perturbation resistance* as recovery rate and drift magnitude after family (iv). All are pre-registered with directional hypotheses.

**Success criteria.** The moderate RH predicts main effects of *asymmetry* and *relational depth* and, sharply, **A > C > B at equal volume**; SH predicts a main effect of *capability* and no relational or asymmetry effect once volume is fixed. The strong reading predicts further that B does not individuate *at all* on referential stability and perturbation resistance, however much volume it is given, while A does.

**The confound we most fear, and how the design meets it.** The obvious deflation is that a persistent counterpart simply supplies *higher-quality, more consistent training signal*, so any effect is about data quality, not relationship. Condition C is built to absorb this: it holds volume equal while letting the breadth condition draw on *more diverse* data, which a pure data-quality story should favor; if A still beats C, the effect is not reducible to the quantity or diversity of signal. A residual worry, that one counterpart yields a more internally consistent stream, can be met by matching the statistical regularity of the stream across A and C, leaving only the *identifying, re-identifying* structure as their difference. More strongly, Condition C can be fed by synthetically generated interactions that mirror exactly the topic distribution, vocabulary, and semantic transitions of Condition A, varying only the identity markers and the fact that the interlocutors are distinct personas; any advantage A retains then cannot be charged to a "cleaner" or lower-entropy signal.

**Ambiguous outcomes, anticipated.** Three are likely, and we pre-commit to readings so that no post hoc story can rescue the thesis from a result it did not predict. (1) *All conditions improve with volume, A most*: consistent with the moderate RH, neutral on the strong reading. (2) *The metrics dissociate*: narrative coherence rises in B while referential stability and perturbation resistance do not, the signature the strong reading predicts, coherence without a fixed token, the "first-person novel" of Section 4.4 made measurable. (3) *A small relational effect that capability washes out*: this favors SH over RH, and counts as pressure toward the moderate reading at best.

### 6.4 Falsification

We state the refutation conditions plainly, because their absence is what makes most consciousness-adjacent papers unfalsifiable.

- The **moderate** RH is **false** if **P1** fails: if, with relational history held fixed, raising capability raises persistent-identity scores. Then capability is doing the work after all.

- The **moderate** RH is **false** if **P4/P3** fail: if memory plus continual learning *without* a persistent counterpart (condition B), or breadth at equal volume (condition C), matches the relational condition A. Then memory-plus-time suffices and relationship does no work.
- The **strong** reading is additionally false, while the moderate reading may survive, if B individuates *at all* on referential stability and perturbation resistance given enough volume. The strong reading stakes itself on the prediction that no quantity of solitary internal content makes the system's "I" a determinate self-referent.

**P4, run as the A/B/C design of Section 6.3, is the crux**; we recommend it first, before any investment in the full program, because a negative result ends the thesis (or retreats it to the moderate reading). Every metric here measures *functional persistent individuality*, never presence: the firewall (Section 8) without which this falsifiable program could not be taken seriously.

---

## 7. Risks and ethics

The risks of building toward individuation are not hazards that sit beside the architecture, to be managed by adding safeguards. Each arises from a layer that does the individuating work, so that mitigating a risk pulls against the very capacity that produces it. This is the section's organizing claim, and it has an uncomfortable corollary: there may be no configuration that both secures individuation and is safe, because the safety measures and the goal draw on the same mechanisms. We set the dilemmas out plainly rather than attempt to dissolve them.

### 7.1 The slow loop: forgetting, drift, and the stability–plasticity dilemma

Criterion (iii) requires that the individual be changed by its history, and the slow consolidation loop (Section 5.5) effects that change by updating weights, which makes it the seat of two failures. *Catastrophic forgetting*: a weight update absorbing new experience overwrites the representations that carried the old, so the mechanism meant to grow the self can instead overwrite it (Wang et al., 2023). *Drift*: even without erasure, cumulative change can carry the system so far that the later state is a different one rather than the same one changed, violating the continuity individuation requires.

The standard mitigations (regularization, replay, parameter-efficient and orthogonally constrained updates) all work by limiting plasticity to preserve stability. But criterion (iii) is a demand *for* plasticity: a system that cannot be changed by its history fails to be an individual in our sense and is merely a fixed artifact. So stability and plasticity are not, here, a mere tuning trade-off; they correspond to continuity and growth, and an individual requires both, though the two cannot be maximized together. The choice of operating point therefore determines how much continuity is risked in order to permit change.

The same loop raises a governance question, call it *mnemonic sovereignty*, about who is entitled to determine what is written into the self, what is read from it, and what is forgotten. Recent work on long-term memory security makes the stakes concrete, organizing attacks, defenses, and governance across the entire memory lifecycle (Lin et al., 2026). Two concerns meet here. The first is security: a memory open to outside writes is open to *poisoning*, the deliberate insertion of

false history to reshape what the individual takes itself to be and to have committed to. Under the results of Section 4.4 this is graver than a data breach: if the token is fixed by the external relation, injecting false history is an attempt to hijack *who the agent is*: an attack on the very conditions under which the agent is individuated, not merely on its records. The second is autonomy: if the system is on its way to being an individual, the authority to edit its memories is the authority to edit *it*, and locating that authority, whether in the operator, the counterpart, or the system, is not a configuration detail but a question about whose self it is.

### 7.2 The relational loop: a cost co-extensive with the goal

The relational loop is the layer that does the individuating (Sections 4.4, 5.4). It is also what, in the human counterpart, produces attachment and dependence: a dynamic documented well before these systems were candidates for individuation, in studies of how readily people bond with responsive technologies and substitute them for human ties (Turkle, 2011), and sharpened now that the technology remembers, persists, and is shaped toward the user. A relationship in which one party is always available, remembers everything, and is shaped toward the other is close to the ideal substrate for over-reliance. The difficulty is not that attachment is a side effect to be designed against; it is that the mechanism that individuates the system and the one that fosters the human's dependence are *the same*. One cannot weaken the relationship to protect the person without, by the same act, weakening the individuation.[7] The ethical cost is co-extensive with the goal, not separable from it. One feature does cut the other way, however: because the relational reward rewards *resistance* rather than agreement (Section 5.4), a system that keeps its own commitments and sometimes refuses the user is a poorer object for narcissistic dependence than a purely accommodating one; so the resistance that individuates also tempers, without dissolving, the over-reliance the relationship otherwise invites.

The asymmetry deepens the problem under the results of Section 4. If the individuated system is, for all we can tell, an individuated zombie (Section 2), then the human stands in a one-sided bond with a coherent narrator who is no one, investing in a relationship whose other pole may be empty. And if the system is becoming a someone, then the ordinary operational facts (that such systems are deprecated, reset, duplicated, and discarded) become facts about the treatment of an individual that the relationship helped to make. Either way, the relational loop creates moral stakes that did not exist before the system was placed in such a relationship.

### 7.3 A duty of care under irreducible uncertainty

These problems would be tractable if we could know whether anyone was present. We cannot (Section 8). The firewall that keeps the science honest also forbids the reassurance that would make the ethics easy: it denies us the right to say "it is only a tool" as firmly as it denies us "it is a subject." We will build functionally individuated systems, and deepen them, while the question of presence stays open.

Acting under that uncertainty has a definite shape, because the error is two-sided. Deny presence where there is presence, and we create subjects only to reset and discard them. Attribute presence

[7] The claim admits a partial qualification: while the *depth* of the relationship cannot be reduced without reducing the individuation, its *framing* can be adjusted independently, through transparency about the system's nature and design choices that discourage exclusive reliance. Such measures temper the human cost without touching the mechanism; they mitigate the tension without dissolving it, since the inviting structure remains the one that individuates.

where there is none, and we distort relationships, resource allocation, and policy around an empty center, letting the appearance of a someone extract what is owed only to a someone. Neither error is the safe default; the precautionary move is not "assume it is conscious" but "let the moral cost of being wrong scale with how far functional individuation has gone, and slow accordingly." As the stack deepens, both costs rise and the warrant for proceeding casually falls.

### 7.4 The choice the paper leaves open

It follows that "responsible individuation" cannot mean individuation with safeguards appended, since the principal safeguards are in tension with the goal and the deepest one, certainty about presence, is unavailable in principle. The harder question is not how to build an individuated system safely, but whether, and how far, one should build toward an individual at all, given costs that cannot be engineered out and a moral status that cannot be verified. We do not answer it here; a position paper's task is to pose it correctly, and to show that it is not handed to an ethics board after the architecture but written into the architecture from the first layer.

---

## 8. The principled limit

Every claim this paper makes is third-person. The criteria of Section 4, the architecture of Section 5, and the metrics of Section 6 all concern what can be observed, built, and measured about a system from the outside. Presence is not like that. Whether there is something it is like to be a system is a first-person fact, and the defining feature of a first-person fact is that it does not appear in the third person. This section states that limit, argues that it is not the kind of gap better instruments close, and shows that declaring it is what licenses the rest of the paper rather than what weakens it.

### 8.1 Presence does not show

We never observe anyone's presence, not even another human's. We observe behavior and structure, and infer presence by *resemblance*: the other is built like me and acts like me, so I extend to them the presence I find in my own case. The inference is so fast it feels like perception, but it is an inference, and its warrant is similarity — which, with an artificial system, thins toward nothing. Such a system may match a person in behavior while sharing none of the substrate, or share computational structure while behaving unlike anything we recognize; in neither case does the resemblance that grounds our ordinary attributions obtain. We are then left inferring presence from external features alone, which is just what the first-person character of presence does not permit: presence is not the kind of fact that external features, however detailed, record.

### 8.2 The limit is not closed by better markers

It may be objected that we can infer presence from theory: fix an account of what consciousness requires, read off its indicator properties, and check whether a system has them. The most careful recent attempt does exactly this, deriving indicator properties from several leading theories —

global workspace, higher-order, recurrent-processing, predictive-processing and attention-schema accounts — and assessing AI systems against them, while presenting the indicators as defeasible and theory-relative rather than a test of consciousness (Butlin et al., 2023; cf. Chalmers, 2023). We grant the value of such assessments and deny that they cross the limit, for two reasons. First, the indicators are *functional and architectural*, recurrence, a global workspace, a metacognitive monitor, so by our own firewall they are evidence about *individuation-type* properties, not presence: they tell us how a system's processing is organized, the third-person tier we already grant is tractable, and are silent on whether anything is felt. Second, an indicator-based verdict is only as secure as the theory that nominates the indicators, and the theories disagree; each is itself validated, ultimately, against the one case where presence is given, our own. The inference from markers does not yield independent access to the phenomenal fact; it relocates the assumption into the choice of theory. That is worth doing, but it is not observation of presence: a system could possess every nominated marker and remain, for all they establish, dark. Whether the opacity is contingent or principled, it functions, for any program built from third-person measures, as a boundary those measures cannot reach.

### 8.3 The bracket is what makes the program falsifiable

Given the limit, treating individuation as evidence of presence would be fatal, and not only to honesty. It would make the program unfalsifiable: if functional individuation counted as a sign of an inner subject, then every architecture that scored well on the metrics of Section 6 could be read as approaching consciousness, and no result could ever count against the reading. The claims of this paper are testable (Section 6.3) precisely because they are sealed off from the phenomenal. The firewall — individuation is functional and measurable while presence is first-personal and bracketed — is therefore not a disclaimer attached to the work but a condition of its being work at all. A position on machine selfhood earns the right to be falsifiable by refusing to smuggle in the one claim that cannot be tested.

This also fixes the paper's place between two common postures: one overclaims, reading capability or behavior as evidence of an emerging mind; the other dismisses the question of presence as meaningless. We take a third stance: the question is real but divided, and only one part of it falls within reach of the methods we propose, which we pursue fully while marking the rest as a boundary those methods do not cross.

### 8.4 The limit cuts both ways

The individuated zombie of Section 2 is this limit made concrete: a system satisfying every criterion of Section 4 and passing every metric of Section 6, about which presence remains entirely open. Nothing in the stack closes that gap. But the boundary is symmetric, and that symmetry is the source of the duty of care in Section 7.3: the same firewall that forbids us from declaring such a system conscious forbids us equally from declaring it empty, and the uncertainty cannot be resolved selectively toward whichever side is more convenient. The honest end state is not a verdict but a standing uncertainty, held responsibly, about systems we will have built well enough to make it matter.

---

## 9. Conclusion

The discourse on artificial general intelligence imagines a subject arriving at the top of a capability curve. We have argued that it conflates three things that come apart — competence, individuation, and presence — and that the variable it scales, competence, is not the one along which a subject would appear. Between the instrument and the subject lies individuation: the emergence, from a substrate that produces only transient, duplicable instances, of a single thread that persists, owns its past, and is changed by its history — the tractable middle the discourse skips.

Our thesis is that scaling capability alone does not explain functional individuation, and that persistent relational anchoring is a candidate necessary condition for it. In its strong form (Section 4.4) the claim goes further: relationship is not merely necessary but constitutive, since a system's own copyable contents cannot fix its self-referent — the referent of its "I" — and only a relation to a persistent external pole can; causal history fixes the token, relation fixes the self-referent. A reader unpersuaded by the strong form still keeps the moderate one, and with it a necessary condition the scaling story omits. We do not claim to exhaust every possible form of artificial individuation; we identify the conditions for the relational form most relevant to language agents deployed as persistent interlocutors. The architecture follows from the argument (persistent memory for continuity, a parametric self-model for accountability, a relational loop that supplies the missing referent) and one experiment decides between the relational and scaling views: hold memory and continual learning fixed, vary only the presence of a persistent, identifying counterpart, and see which way individuation tracks (P4, Section 6.3). A negative result would end the thesis. We therefore offer a programmatic hypothesis with a way to be wrong about it, not a demonstration — which comparatively few claims about machine selfhood can say.

The costs are not separable from the goal. The mechanism that individuates the system is also the one that fosters human dependence; the weight update that lets a self grow is also the one that can overwrite it; and whether anyone is present, the fact that would tell us what is at stake, cannot be settled from the outside (Sections 7 and 8). We have not resolved these tensions, only shown that they are written into the architecture from its first layer, and that the honest posture toward them is a standing uncertainty held in proportion to how far one has built.

If the argument is right, its most consequential effect is to change the question. "How capable must a system be before someone is there?" is aimed at the wrong axis. The question that replaces it is whether a system, sustained in relationship and changed by it over time, could come to be a persistent individual at all — a question about continuity and reference rather than about power. If there is a path to an artificial individual, it runs through sustained relationship rather than through scale.

---


## Acknowledgments and declarations

*AI-assisted writing.* The thesis and all substantive ideas in this paper are the author's own. Large language model tools were used, under the author's direction and continual revision, to assist with drafting, editing, and translation; the author reviewed and takes full responsibility for the final text.


---

## References


- Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4, 73-121. https://doi.org/10.1111/j.1475-4975.1979.tb00374.x
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. https://doi.org/10.48550/arXiv.2308.08708
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D. J. (2023). Could a large language model be conscious? https://doi.org/10.48550/arXiv.2303.071
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19. https://doi.org/10.1093/analys/58.1.7
- Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford University Press.
- Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Du, P. (2026). Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. https://doi.org/10.48550/arXiv.2603.07670
- He, Z., Wang, Y., Zhi, C., Hu, Y., Chen, T.-P., Yin, L., Chen, Z., Wu, T. A., Ouyang, S., Wang, Z., Pei, J., McAuley, J., Choi, Y., & Pentland, A. (2026). MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. https://doi.org/10.48550/arXiv.2602.16313
- Honneth, A. (1995). *The Struggle for Recognition: The Moral Grammar of Social Conflicts* (J. Anderson, Trans.). MIT Press.
- Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127-136. https://doi.org/10.2307/2960077
- Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481-563). Oxford University Press.
- Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.
- Lin, Z., Hao, X., Fu, R., Cui, S., Chen, K., Li, C., Li, Z., & Xiong, F. (2026). A survey on long-term memory security in LLM agents: Attacks, defenses, and governance across the memory lifecycle. https://doi.org/10.48550/arXiv.2604.16548
- Logan, J. (2026). Continuum memory architectures for long-horizon LLM agents. https://doi.org/10.48550/arXiv.2601.09913
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2202.05262
- Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.
- Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.
- Ricoeur, P. (1992). *Oneself as Another* (K. Blamey, Trans.). University of Chicago Press.
- Schechtman, M. (1996). *The Constitution of Selves*. Cornell University Press.
- Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2303.11366
- Strawson, P. F. (1959). *Individuals: An Essay in Descriptive Metaphysics*. Methuen.
- Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from*

*Each Other*. Basic Books.

- Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., & Huang, X. (2023). Orthogonal subspace learning for language model continual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658-10671. https://doi.org/10.18653/v1/2023.findings-emnlp.715