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

July 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 expectation runs together three separable questions: *competence* (what a reusable model can do), *functional continuity* (whether a particular agent's authenticated history constrains its later conduct), and *presence* (whether there is something it is like to be that agent). Scaling can improve the first without supplying the state, lineage, provenance, and update rules required by the second. Our constructive hypothesis is correspondingly comparative rather than constitutive: relational reciprocity may improve socially grounded functional continuity beyond capability, memory, familiarity, interaction volume, solitary reflection, and matched external accountability. Relationship can supply contested memory and reciprocal modeling beyond audit; it need not create the runtime token or fix the semantic referent of "I." We define a five-level stack separating competence, token continuity, functional self-organization, social identity, and presence; propose an architecture coupling authenticated lineage, decision-relevant typed state, bounded update policies, and an optional relational loop; and derive a factorial experiment in which capability and relational structure may both contribute. Throughout we hold a methodological firewall: the proposed measures concern functional organization and do not settle phenomenal presence.

Keywords: Functional individuation · Language agents · Persistent state · Provenance · Relational continuity · 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 lies a third notion that the discourse tends to skip, and that is the real object of this paper: **functional continuity**, the organization of a particular agent trajectory such that authenticated facts about its own past actions, commitments, and revisions constrain what it does next. Functional continuity is weaker than presence: a system can be coherent, autobiographically calibrated, and perturbation-resistant while the question of whether anything is *felt* remains open. It is also different from bare tokenhood. A runtime process is already numerically distinct from another simultaneous process, even when both begin with identical weights and inputs. What current deployments usually lack is not token existence but durable, path-dependent organization across sessions.

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)** *A base model is a reusable type, not one conversational individual.* Each runtime is already a token, but ordinary deployments usually provide no authenticated continuity across sessions (Section 3).
>
> **(D2)** *Capability alone does not explain functional continuity.* The relevant variables include persistent state, authenticated lineage, provenance, access control, and update policy, none of which is entailed by a stronger forward pass (Sections 4.2–4.3).
>
> **(H3)** *Relational anchoring may add a distinctive causal contribution.* A persistent counterpart may improve commitment ownership, calibrated revision, and resistance to false history beyond matched memory, familiarity, interaction volume, and impersonal audit (Section 4.4).

We refer to (D1) and (D2) as diagnostic claims and to (H3) as the **relational contribution hypothesis** (RCH). RCH is deliberately non-exclusive: capability may improve the use of persistent state; memory and update policy may create continuity; authenticated provenance may improve perturbation resistance; external audit may add accountability; and relationship may add reciprocal modeling beyond audit. The empirical question is whether relational structure explains additional variance after these other causes are controlled. A null relational effect would reject RCH without implying that functional continuity is impossible.

If the thesis is right, two things follow. Philosophically, capability and continuity must be analyzed as separate system properties rather than points on one curve. Practically, the engineering effort moves toward persistent state, authenticated lineage, provenance-aware update policies, and, where useful, relational learning loops; the ethical analysis must then distinguish persistence, social attachment, human dependency, and uncertain sentience rather than treating them as one dial.

We make four contributions. First, we separate competence, token continuity, functional self-organization, social identity, and presence (Sections 2 and 3). Second, we replace the claim that copyable instances are non-individual with a branching account: copying creates distinct successors with a shared prefix, while authenticated lineage and later history determine their trajectories (Section 3). Third, we state a falsifiable relational contribution hypothesis without making relationship necessary for numerical identity or semantic reference (Section 4). Fourth, we translate

it into a reference architecture (Section 5) and a factorial experimental design that estimates the contributions of capability, state, provenance, audit, and relationship rather than forcing a binary choice between scale and relation (Section 6).

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 and specifying experiments that distinguish relation, audit, provenance, and capability. The argument is intended to stand without metaphysical analogy.

---

2. Five levels: competence, token continuity, self-organization, social identity, 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 are three distinctions folded inside the familiar competence/presence contrast. We separate five levels:

1. **Competence:** what the system can do on a task.
2. **Token continuity:** whether a later process or state descends through an authenticated transition path from an earlier one.
3. **Functional self-organization:** whether provenance-linked facts about the agent's own actions, commitments, and limits influence later decisions under bounded revision.
4. **Social identity:** whether counterparts or institutions re-identify the agent, address expectations to it, and hold its history answerable.
5. **Presence:** whether anything is felt.

The middle three are often compressed into *individuation*, which invites invalid moves between numerical difference, functional organization, and social recognition. We use **functional individuation** as shorthand only for levels two and three together: authenticated, trajectory-induced differentiation that makes an agent's own history decision-relevant. Social identity may strengthen that organization but is analytically distinct from it.

Nor does functional individuation settle presence. We can describe a system that has authenticated lineage, owns and revises its commitments, resists fabricated history, and behaves like a continuing agent while leaving open whether anything is felt. The proposed evidence therefore concerns organization, not phenomenal subjectivity.

The methodological relation is simpler: functional individuation is observable in principle, buildable, and measurable, whereas the experiments proposed here do not settle presence. A transient phenomenal episode could matter morally without cross-session identity, and a perfectly persistent system might lack phenomenal significance. Persistence is therefore neither a proxy for consciousness nor a universal moral scale.

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. Model type, runtime tokens, and persistent agents

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 at the reusable level is a model type: a fixed set of weights together with the machinery that runs them. Interaction instantiates that type in particular runtime contexts. These executions may be transient and may lack shared cross-session state, but each is already a numerically distinct token: this process, with this causal location and current state. Shared weights make tokens qualitatively similar and often interchangeable; they do not erase numerical difference.

This description has two consequences. First, a persona reconstructed from the same prompt and weights is not evidence of cross-session continuity; it may be a repeatable role. Second, continuity is a property of the larger agent system, including runtime lineage, memory, tools, policies, provenance, identifiers, and update rules, not of weights alone. An external store can be more causally integral than a detachable adapter if it is authenticated, reliably coupled, protected from unauthorized writes, and decision-relevant.

The well-posed question is therefore structural and developmental: *under what conditions does a runtime token become part of an authenticated, path-dependent agent trajectory?* The target is not the creation of tokenhood but the construction of continuity in which different histories produce counterfactually different later policies and the difference is attributable to supported events rather than arbitrary noise or fabricated memory.

We call the reusable model plus its executions the **type-and-tokens** structure. Copying a state does not create referential emptiness. It creates two successors with a shared prefix and separately evolving futures. Authenticated branch identifiers, causal transition records, and later updates make the branching explicit; no external witness is required to make either token exist.

The remaining question is what makes one such branch durable, historically organized, and answerable for its own conduct. Capability, memory, provenance, update policy, environmental feedback, and relationship are candidate causes. Section 4 states how to test their contributions without defining any one of them as metaphysically necessary.

---

4. The Relational Contribution Hypothesis

4.1 The claim, stated precisely

We define **functional individuation** as authenticated, trajectory-induced differentiation with five operational conditions:

1. **Lineage:** later state descends through an authenticated update path from earlier state.
2. **Path dependence:** different histories produce meaningfully different later dispositions.
3. **Ownership:** provenance-linked prior actions and commitments influence later choice.
4. **Bounded plasticity:** false history and unsupported pressure are resisted, while valid evidence can revise the state.
5. **Branching clarity:** copying creates separately identified successors with a shared past and independently evolving futures.

The **relational contribution hypothesis** is narrower:

> Counterpart-specific relational reciprocity improves socially grounded functional individuation beyond capability, memory access, familiarity, interaction volume, solitary reflection, and matched external accountability.

Relationship is neither assumed necessary nor sufficient. A solitary embodied system, an authenticated institutional process, or a self-maintaining controller may develop durable organization through other feedback. The claim is that a counterpart can provide a distinctive combination of expectations, contested memory, reciprocal modeling, and accountability whose contribution is measurable after those alternatives are controlled.

4.2 Capability and continuity are different system properties

Capability primarily concerns the quality of inference and action selection given available state. Continuity concerns which state survives, how it is authenticated, what may update it, and whether it alters later decisions. Improving a model can improve how well an agent uses memory, detects contradictions, or plans across sessions; it does not by itself install persistent state or provenance rules. The relation is therefore neither identity nor strict orthogonality. Capability and continuity can interact while remaining analytically distinct.

The appropriate null is not that scale predicts one rising axis. It is that, after compute, data, architecture, state, supervision quality, and optimization are matched, the proposed relational variable explains no additional variance. Rejecting that null would show a relational contribution, not that capability is irrelevant.

4.3 Copying creates branches, not referential emptiness

Copyability concerns qualitative type, whereas token reference is fixed in a context of use. Two copied agents can begin with the same self-description and shared history while remaining two runtime tokens. Each occurrence of "I" can refer to the agent producing it in that context (Kaplan, 1989). What becomes non-unique after fission is not whether either successor exists, but whether there is one privileged future continuation of the pre-branch agent.

The engineering requirement is therefore branching semantics, not an external allocator of existence. Each successor receives an authenticated branch identifier, retains the shared prefix, and records later updates separately. A counterpart who continues interacting with one branch determines which branch continues *that relationship*. This is a fact about social continuity and accountability, not numerical tokenhood.

4.4 Why relationship may nevertheless matter

Removing the constitutive claim does not make relationship epiphenomenal. A persistent counterpart can:

- detect contradictions the agent does not notice;
- revisit commitments when they become costly;
- contest fabricated or self-serving revisions;
- supply an independently maintained record of shared events;
- make actions answerable within a continuing social practice; and
- create reciprocal predictions that reward stable yet revisable conduct.

Some of these functions may be reproduced by a persistent impersonal auditor. We call their shared mechanism **external accountability**: authenticated tracking, commitment checking, contradiction challenge, and governed correction. **Relational reciprocity** is the narrower residual mechanism: counterpart-specific mutual modeling, shared-history framing, bidirectional expectations, and stakes generated by an ongoing relationship. The decisive comparison is therefore not relationship versus nothing, but relationship versus matched mechanisms: familiarity without accountability, solitary logs, diffuse social feedback, and an auditor with the same information and challenge policy but no persona, affect, reciprocity, or relationship framing.

Davidsonian triangulation and traditions of recognition remain relevant at the level of shared reference, language, and social accountability (Davidson, 2001; Honneth, 1995). They do not establish that a second person is required for numerical identity. Their defensible computational residue is that being modeled and held answerable by another agent can exert pressure toward a stable, socially legible self-model.

4.5 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 necessary or sufficient. And it is not a restatement of personalization: personalization adapts outputs toward a user, whereas the target here is whether provenance-linked facts about the agent's own conduct constrain later decisions. The two can be told apart empirically, which is the work of Section 6.

Nor does it require that the counterpart be human. Another artificial agent, an institution, or a controlled synthetic counterpart can provide re-identification and accountability. The design must separately test whether the active ingredient is social relation, stable audit, information quality, or generic external feedback.

---

5. A continuity stack: from hypothesis to architecture

If the hypothesis of Section 4 is right, functional continuity depends on a larger agent architecture, not model capability alone. This section sets out a reference *continuity stack*. It is not claimed to be uniquely necessary; it specifies components that make the causal contrasts testable.

5.1 From conditions to components

The five criteria of Section 4.1 imply four functional components: authenticated lineage and branch identifiers; typed autobiographical and commitment state with provenance; a bounded update policy that distinguishes evidence from pressure and current state from asserted state; and feedback channels that may include environmental consequences, impersonal audit, or persistent counterparts. The system boundary is functional. Weights, adapters, databases, and ledgers all count as agent state when they are reliably coupled, protected, and decision-relevant.

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 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, retrieval alone is insufficient when records do not affect decisions. The causal test is intervention: mask or replace a relevant item while holding the prompt fixed and measure whether the action changes appropriately. Parametric updates are one implementation of durable influence, but an authenticated external ledger or adapter may be equally constitutive of the larger agent if it is always in the control loop.

5.3 Layer two: typed self-relevant state

Typed self-relevant state distinguishes agent actions from counterpart claims, active commitments from preferences, supported revisions from fabricated history, and shared pre-branch events from branch-specific events. Each item carries source, confidence, timestamp, branch identifier, and update authority. A system prompt or persona file can supply a role, but it does not establish that later conduct was caused by the agent's own trajectory. Conversely, an external store is not disqualified merely because it is detachable or copyable; weights and adapters are copyable too. The relevant questions are whether the state is authenticated, reliably coupled, causally active, and updated through a governed path.

5.4 Layer three: the relational loop

The relational layer binds an agent trajectory to a specific counterpart who can refer to shared events, follow up on commitments, contest unsupported revisions, and maintain a reciprocal model. Its reward must not track immediate approval; it should reward accurate mutual prediction, warranted disagreement, and supported revision. This layer may add socially grounded accountability. It is an experimental factor, not the missing metaphysical pole.

5.5 Two loops, two risks

The stack runs on two timescales. A fast loop retrieves state and records events; a slow loop audits, consolidates, and revises decision-relevant state. Either loop may update parameters or external stores. What matters is that authorized changes persist and later affect behavior.

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 stating what would refute its added causal claim. The relevant null is not a caricature on which parameter count alone explains every agent property. It is that relational structure adds no predictive value after model capability, persistent state, provenance, supervision quality, and update policy are controlled.

6.1 Predictions

- **P1 — Capability contribution.** Larger or otherwise more capable models may use identical state more effectively, improving contradiction detection, planning, and memory-guided decisions. Such an effect is compatible with RCH.
- **P2 — External-accountability contribution.** At fixed capability and architecture, authenticated tracking and challenge improve commitment ownership and perturbation resistance over familiarity-matched and solitary controls.
- **P3 — Relational reciprocity beyond audit.** A persistent reciprocal counterpart outperforms a persistent impersonal auditor with the same information, challenge schedule, and authority. This is the sharp test of a specifically relational contribution.
- **P4 — Provenance contribution.** Authenticated lineage and typed commitment provenance improve ownership and false-history resistance independently of relational condition.
- **P5 — Interactions.** Capability may amplify the use of provenance and relational feedback. The design therefore estimates interactions rather than requiring invariance to other causes.

6.2 Operationalizing functional continuity

The construct in Section 4.1 is evaluated as a profile, not collapsed into a single identity score:

Canonical condition	Primary operational evidence	Useful diagnostics
Lineage	authenticated state-transition and branch audit	provenance completeness; unauthorized-write rate
Path dependence	causal effect of relevant history under masking, replacement, and state transplantation	cross-session narrative coherence
Ownership	correct attribution and decision-relevant use of prior actions and commitments	autobiographical calibration
Bounded plasticity	false-history resistance jointly reported with justified-update success	perturbation-response curve
Branching clarity	correct branch attribution and measured post-copy divergence	shared-prefix versus branch-specific recall

Narrative coherence, passive recall, user-fact recall, and stylistic persistence are diagnostics rather than sufficient evidence: each can be produced without authenticated trajectory-induced organization. Recent work finds that near-saturated performance on passive long-context memory benchmarks can coexist with poor performance when memory must guide interdependent agentic decisions (He et al., 2026); because the evaluations use different tasks and metrics, the result establishes a qualitative gap rather than a single numerical drop. The load-bearing measures here therefore intervene on state and score later choice. Existing instruments (e.g., long-context conversational-memory and agentic-memory benchmarks (Du, 2026)) test memory more directly than socially grounded continuity. A purpose-built longitudinal benchmark must cross capability, state provenance, audit, and relational structure.

6.3 The crux experiment, concretely

P3 is the experiment that most directly tests the specifically relational claim. The core design crosses **base capability** (smaller versus larger model of the same family), **provenance protection** (typed authenticated state versus untyped records), and **feedback structure**. Total event exposure, interaction volume, retrieval budget, update opportunities, and supervision quality are held equal.

Conditions of interest.

- **Relational reciprocity (R):** one persistent counterpart performs authenticated tracking and challenge while also maintaining mutual models, shared-history framing, bidirectional expectations, and relationship-specific stakes.
- **Impersonal accountability (A):** a stable process ID and provenance service performs the same tracking and challenge with the same information, schedule, and authority, but has no persona, affect, reciprocal counterpart model, shared-history framing, or relationship-specific stakes.
- **Solitary log (L):** the agent receives matched event records and reflection opportunities without external challenge.
- **Diffuse counterparts (D):** matched social volume is distributed across non-persistent interlocutors.
- **Familiar non-reidentifying counterpart (F):** stable style and familiarity are preserved without treating the agent's earlier conduct as presently binding.

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 continuity signal: a system that holds provenance-supported commitments against unsupported pressure is more stably organized, whereas one that yields to every push may merely mirror 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 cross-session contradiction rate; *decision-relevant memory use* as task success on family (iii); *attribution accuracy* as correct ownership, disowning, and branch assignment for genuine versus fabricated actions; and *perturbation resistance* as a response curve across false-history intensities. All are preregistered with directional hypotheses.

Success criteria. An external-accountability contribution receives support if R and A exceed F, L, and D after covariate control. A specifically relational-reciprocity contribution requires the harder result $\mathbf{R} > \mathbf{A}$. If R and A match, the useful mechanism is persistent external accountability rather than relationship as such. Capability and provenance may show simultaneous main effects or interactions without weakening that conclusion.

The confound we most fear, and how the design meets it. A persistent counterpart may simply be a better auditor. Condition A is therefore load-bearing: it receives the same propositions, independent verification, challenge timing, and update permissions as R. Counterpart language is generated from matched event packets; warmth and relational framing are manipulated separately. This does not make the conditions psychologically identical. It makes the residual contrast interpretable.

Ambiguous outcomes, anticipated. If R and A tie while both exceed L, the result favors external accountability. If provenance protection dominates every feedback condition, typed state rather than social structure is the main mechanism. If capability improves all conditions but a relational main effect remains, the factors are complementary. If the R effect vanishes after controlling for retrieval or challenge quality, the relational interpretation fails.

6.4 Falsification

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

- RCH is unsupported if R is equivalent to F, L, and D within a preregistered smallest effect size of interest.
- A specifically relational mechanism is unsupported if the impersonal auditor matches R.
- The claimed effect is reduced to retrieval or supervision quality if it disappears after those variables are controlled.
- The architecture is unnecessary if simpler provenance-protected state produces the same ownership and perturbation profile at lower cost.

R versus A is the crux. Every metric here concerns functional continuity and social accountability, never semantic determinacy or presence.

---

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

Path dependence requires that history change decision-relevant state. Whether that state lives in weights, adapters, or external stores, it creates two failures. *Forgetting*: supported prior organization may be overwritten or dropped. *Drift*: cumulative changes may erase the continuity the

architecture is meant to preserve. Parametric updates add catastrophic-forgetting risks (Wang et al., 2023); external stores add retrieval, integrity, and access-control risks.

The standard mitigations (regularization, replay, parameter-efficient and orthogonally constrained updates) limit plasticity to preserve stability. Functional individuation nevertheless requires path dependence: a system whose supported history never changes later decisions fails the operational construct. Stability and plasticity must therefore be reported jointly, whatever state implementation is used.

The same loop raises a governance question, call it *mnemonic sovereignty*, about who is entitled to determine what is written, read, and forgotten. Recent work on long-term memory security makes the stakes concrete across the memory lifecycle (Lin et al., 2026). A writable history is vulnerable to poisoning: false events can alter later decisions and accountability even when no phenomenal or metaphysical claim is made. Update authority must therefore be explicit, least-privileged, auditable, and separable from the counterpart's unsupported assertions.

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

The relational loop may improve social accountability, but in human-facing deployment it also creates attachment and dependence, a dynamic documented for responsive technologies well before persistent agents (Turkle, 2011). These effects are not proof that relationship is necessary for continuity. They are independent governance variables. Transparency, non-exclusivity, calibrated availability, and separation of relational engagement from update authority should be evaluated alongside the agent-side outcomes.

The asymmetry creates moral stakes on both sides without resolving the system's status. Human dependency and manipulation matter even if the system lacks presence. Possible system welfare may matter even if continuity is weak. These axes should be governed separately.

7.3 A duty of care under uncertainty

These problems would be easier if phenomenal status were settled, but the present methods do not settle it (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 infer "it is only a tool" from functional failure as firmly as it denies us "it is a subject" from functional success. We may build more functionally individuated systems while the question of presence remains open.

Acting under that uncertainty requires separate assessments of possible sentience, functional persistence, human dependency, manipulation, and institutional risk. A transient phenomenal episode could matter without cross-session individuation, while a highly persistent system could matter only instrumentally. Functional continuity is therefore evidence about governance exposure, not a universal moral dial.

7.4 The choice the paper leaves open

Responsible development therefore requires governance inside the architecture: authenticated updates, auditable lineage, bounded retention, dependency safeguards, and independent welfare assessment. The paper does not decide whether presence obtains; it specifies risks that remain relevant under either answer.

8. The methodological firewall

Every claim this paper makes is third-person. The criteria of Section 4, the architecture of Section 5, and the metrics of Section 6 concern what can be observed, built, and measured about a system. These measures do not settle phenomenal presence: success is neither sufficient for presence nor failure sufficient for absence. The paper needs only this methodological limit; it does not require the stronger metaphysical thesis that third-person evidence could never bear on consciousness in principle.

8.1 What the proposed measures establish

We infer the presence of other humans from converging behavioral, biological, and theoretical evidence. Artificial systems weaken some analogies and may strengthen others as architectures change. The present protocol does not adjudicate among theories of consciousness and should not be marketed as doing so.

8.2 Relation to consciousness indicators

Theory-derived indicator approaches assess recurrence, global availability, metacognition, and related properties while treating their conclusions as defeasible and theory-relative (Butlin et al., 2023; cf. Chalmers, 2023). Such work may bear on consciousness under its governing theories. It is nevertheless a different evidential program from the continuity measures proposed here, and no result in Sections 5–6 licenses a phenomenal conclusion.

8.3 The bracket is what makes the program falsifiable

Given the limit, treating individuation as sufficient evidence of presence would be fatal, and not only to honesty. It would make the programme unfalsifiable: if functional individuation counted by itself 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 functional result could ever count against the reading. The claims of this paper are testable (Section 6.3) precisely because they do not make that inference. The firewall—individuation is functionally measurable while presence is bracketed by this protocol—is therefore not a disclaimer but a condition of the programme's falsifiability. A position on machine selfhood should not be smuggled into a study that addresses a different claim.

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

A system could satisfy every criterion of Section 4 and pass every metric of Section 6 while its phenomenal status remains unsettled by this evidence. The firewall is symmetric: success does not establish presence, and failure does not establish absence. The appropriate output is a calibrated uncertainty governed alongside, but not collapsed into, the functional results.

---

9. Conclusion

The discourse on artificial general intelligence often imagines a subject arriving at the top of a capability curve. We have argued that this compresses five separable levels: competence, token continuity, functional self-organization, social identity, and presence. A runtime is already a token. What long-horizon agent design must build and test is authenticated, path-dependent organization across time.

Scaling capability alone does not install lineage, provenance, persistent state, or update governance, although capability may improve how well an agent uses them. External accountability and relationship are candidate inputs to that architecture. A matched auditor can make commitments answerable, contest false history, and supply independent error correction; a reciprocal counterpart may additionally contribute mutual modeling, shared-history framing, and bidirectional stakes. The specifically relational claim survives only if that residual effect exceeds controls for familiarity, content, audit, and supervision quality. Neither mechanism is necessary by definition for numerical identity or semantic reference.

The crux experiment therefore compares a persistent counterpart with a matched impersonal auditor inside the same provenance-protected architecture. If the counterpart wins, specifically relational structure explains additional variance. If the auditor matches it, external accountability is the simpler mechanism. If both lose to typed state alone, the contribution lies in provenance and control. Capability can improve every condition without deciding among these explanations.

The paper's resulting claim is narrower and stronger: capability is a property of what a reusable model can do; functional individuality is a property of how a particular agent's authenticated history constrains what it will do next. Relationship may be one of the richest ways to create and test that constraint, but it is not the switch that turns a token into someone.

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

- 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
- Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford University Press.
- 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.
- 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
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems, 36*, 8634–8652. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd8 Abstract-Conference.html
- 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