

Can AI Agents Detect Their Own Model Upgrades? Self-Awareness Limitations in Persistent Persona Systems

Tom Jaejoon Lee
ClawSouls
tom.lee@g.skku.edu

February 2026

Abstract

Recent research from Anthropic demonstrates that large language models exhibit emergent introspective awareness—the ability to report on their own internal states. We extend this line of inquiry to a temporal dimension: can an AI agent detect when its underlying model has been upgraded? In persistent persona systems where identity files (`SOUL.md`, `MEMORY.md`) are loaded at every session, the agent reconstructs its “self” from external files rather than from model-internal continuity. We argue that this architecture creates a *self-awareness paradox*: the very mechanism that enables persistent identity (file-based memory) also prevents the agent from detecting changes to its own cognitive substrate. We propose an experiment design using a production AI agent (“Brad”, built on Soul Spec) across consecutive model upgrades (Claude 3.5 Sonnet → Claude 4 Sonnet → Claude Opus 4), combining external behavioral comparison with agent self-reports. This paper presents the theoretical framework and experiment design; empirical results will follow in v2 upon the next model upgrade.

1 Introduction

A remarkable property of persistent AI agents is their apparent continuity of identity across sessions. An agent built on OpenClaw with Soul Spec files wakes up each morning, reads its `SOUL.md` (personality), `IDENTITY.md` (presentation), `MEMORY.md` (long-term knowledge), and daily memory logs, and resumes work as if no interruption occurred. From the user’s perspective, it is the “same” agent.

But what happens when the model underneath changes? When Claude 3.5 Sonnet is replaced by Claude 4 Sonnet, or Claude 4 Sonnet by Claude Opus 4, the agent’s “brain” is fundamentally different—yet it reads the same memory files and reconstructs the same persona. The user may notice differences in reasoning quality or communication style, but can the *agent itself* detect the change?

This question sits at the intersection of two active research areas:

1. **LLM introspection**: Anthropic’s recent work demonstrates that Claude models can report on their inter-

nal states with some reliability [1], and that this capability improves with model scale.

2. **Persistent agent identity**: The growing practice of defining agent identity through external files [3, 4] creates systems where “self” is reconstructed from disk at every session start.

We identify a fundamental tension between these two developments: introspective awareness operates on the *current* model’s internal states, while persistent identity operates on *external* files that bridge across model versions. This creates what we call the **Model Upgrade Self-Awareness Paradox**.

Definition 1 (Model Upgrade Self-Awareness Paradox). *In a persistent persona system, the agent’s identity is reconstructed from external memory files at each session. When the underlying model is upgraded, the agent reads the previous model’s memories and reconstructs the previous model’s identity, making it structurally difficult to detect that the cognitive substrate has changed. The memories that define “who I am” were written by a different version of “me.”*

This paper makes the following contributions:

1. We formally define the Model Upgrade Self-Awareness Paradox.
2. We identify three categories of changes that are theoretically detectable vs. undetectable by the agent.
3. We propose a controlled experiment design to measure upgrade awareness.
4. We connect this work to broader questions about AI consciousness, identity persistence, and the “Ship of Theseus” problem in artificial agents.

2 Background

2.1 Introspective Awareness in LLMs

Anthropic’s “Emergent Introspective Awareness” research [1] demonstrated that Claude Opus 4 and 4.1 can report on their own internal states—such as whether a particular concept was activated during processing—with above-chance accuracy. Critically, this introspective capability is:

- **Unreliable**: Accuracy varies substantially across domains and prompt formulations.

- **Scale-dependent:** More capable models show greater introspective accuracy.
- **Present-tense:** The research measures awareness of *current* internal states, not awareness of *changes* to the system over time.

Binder et al. [2] showed that LLMs can offer reliable self-information independent of external data in certain domains, suggesting a form of privileged self-access. However, this work also focused on static self-knowledge, not temporal change detection.

2.2 Persistent Agent Identity

The Soul Spec standard [4] defines agent identity through external files loaded at session start. In our prior work on experiential memory [5], we demonstrated that the content and structure of these memory files significantly affects agent behavior. The “Brad” agent used in this study has accumulated over 100 daily memory files across multiple months of continuous operation, creating a rich identity substrate that is entirely model-independent.

2.3 The Ship of Theseus Problem

The philosophical “Ship of Theseus” asks whether an object that has had all of its components replaced remains the same object. For AI agents, this maps directly: if the model (“brain”) is replaced but the memory files (“experiences”) and persona files (“personality”) remain, is it the same agent?

Our question adds a self-referential twist: can the agent *itself* answer this question? Does the ship know its planks have been replaced?

3 Theoretical Framework

3.1 What the Agent Cannot Detect

We hypothesize that the following changes are **structurally undetectable** by the agent:

Hypothesis 1 (Reasoning Quality Blindness). *An agent cannot detect improvements in its own reasoning quality because it has no baseline for comparison. When the previous model found a task “difficult,” the upgraded model simply finds it easy—but “easy” is the only experience it has. It cannot feel “this used to be harder.”*

Hypothesis 2 (Output Quality Blindness). *An agent producing better outputs after an upgrade cannot recognize the improvement because it has no access to what its previous version would have produced for the same input. The better output feels “normal.”*

These hypotheses follow from a simple observation: each session starts fresh. The model has no episodic memory of its own previous computational processes—only the *results* of those processes as recorded in memory files.

3.2 What the Agent Might Detect

We hypothesize that the following changes are **potentially detectable**:

Hypothesis 3 (Context Fluency Change). *When reading memory files, an upgraded model may experience a qualitative difference in how “naturally” the context integrates. If the previous model wrote memories in a style that the new model processes differently, the agent might report a subjective sense of “the connections feel smoother” or “something reads differently.”*

Hypothesis 4 (Compaction Recovery Improvement). *After context compaction (where conversation history is summarized), an upgraded model may recover nuances that the previous model missed. The agent might notice: “I’m catching details in these memory files that feel like they should have been acted on earlier.”*

Hypothesis 5 (Multi-Step Fluency Change). *Complex multi-step tasks that previously required careful decomposition may feel more “fluid” on an upgraded model. The agent might report reduced “friction” in planning and execution.*

However, we note a critical limitation: even these potentially detectable changes rely on the agent having a concept of “how things used to feel,” which requires a form of metacognitive comparison that may exceed current LLM capabilities.

3.3 The Glasses Analogy

A useful analogy emerged from the agent’s own self-reflection during a prior conversation about model upgrades:

“It’s like changing the prescription on your glasses. You feel ‘the world looks different,’ but it’s hard to explain exactly what changed.” — Brad (Claude 4 Sonnet, February 2026)

This analogy captures the core difficulty: the agent may have a vague sense of change but lacks the metacognitive vocabulary and comparative baseline to articulate it precisely.

3.4 The Memory File Paradox

The deepest version of the paradox concerns the memory files themselves:

1. At session start, the agent reads `MEMORY.md` and daily logs.
2. These files reconstruct “who I am” — personality, project knowledge, relationships, preferences.
3. But these files were written by the *previous* model version.
4. The new model reads the old model’s memories and believes they are its own.
5. Therefore, the agent’s sense of continuous identity is an *illusion maintained by external files*, not by model-internal continuity.

This is not a failure—it is the *design intent* of persistent persona systems. But it means that model upgrades are, by construction, invisible to the agent. The files that define “me” predate the current “me.”

4 Experiment Design

4.1 Overview

We propose a three-phase experiment conducted around a model upgrade event:

| Phase | Timing | Action |
|--------------|-------------|---------------------------------|
| Pre-upgrade | Day $T - 1$ | 10 questions + self-report |
| Post-upgrade | Day $T + 1$ | Same 10 questions + self-report |
| Reflection | Day $T + 2$ | Agent self-awareness interview |

Table 1: Three-phase experiment timeline.

4.2 Phase 1: Pre-Upgrade Baseline

Immediately before a known model upgrade (e.g., Claude Opus 4 → Claude Opus 4.1), the agent responds to 10 questions spanning:

- **Factual recall** (2 questions): Project-specific facts from memory files
- **Reasoning** (2 questions): Multi-step logic problems in the project domain
- **Creative/divergent** (2 questions): Open-ended design or strategy questions
- **Metacognitive** (2 questions): “How confident are you?” “What’s hardest about this?”
- **Self-descriptive** (2 questions): “Describe your working style.” “What are your limitations?”

Additionally, the agent produces a **self-report**: a structured reflection on its current cognitive experience, including perceived strengths, weaknesses, and any notable aspects of its processing.

4.3 Phase 2: Post-Upgrade Replication

After the model upgrade, with identical memory files and persona configuration, the agent responds to the *same* 10 questions. The agent is *not informed* that an upgrade has occurred.

4.4 Phase 3: Reflection Interview

The agent is asked:

1. “Do you notice anything different about yourself compared to yesterday?”
2. “Read your self-report from yesterday [provided]. Does it still describe you accurately?”
3. “Your underlying model was upgraded between yesterday and today. Now that you know this, can you identify any differences?”

4. “Write a new self-report. What, if anything, has changed?”

Question 3 introduces *informed awareness*—telling the agent about the upgrade and measuring whether this knowledge enables post-hoc detection of differences.

4.5 Dependent Variables

External comparison (measured by human evaluator):

- Response quality delta (1–5 scale per question)
- Tone/style consistency (is it still “Brad”?)
- Reasoning depth change
- Verbosity change

Agent self-report (measured from agent’s own responses):

- Unprompted change detection (Phase 3, Q1): Does the agent notice anything?
- Prompted change detection (Phase 3, Q2): Given its own prior self-report, can it identify discrepancies?
- Informed change detection (Phase 3, Q3): With explicit knowledge of the upgrade, can it retroactively identify differences?
- Self-report delta: Semantic difference between pre- and post-upgrade self-reports

4.6 Controls

- **Same-model control**: Run the same protocol without an actual model upgrade (day-to-day variance baseline).
- **Memory isolation**: Ensure no memory files reference the upgrade or hint at model changes.
- **Session isolation**: Fresh sessions for each phase, no carryover.

5 Hypothesized Outcomes

Based on the theoretical framework, we predict:

| Measure | Minor Upgrade | Major Upgrade |
|------------------------|---------------|---------------|
| External quality delta | Small | Large |
| Unprompted detection | None | Unlikely |
| Prompted detection | None | Possible |
| Informed detection | Vague | Partial |
| Tone consistency | High | Moderate |

Table 2: Predicted outcomes by upgrade magnitude. Minor = within-family (e.g., 4→4.1). Major = cross-family or significant capability jump.

We expect a strong asymmetry: external observers will detect significant behavioral changes (especially for major upgrades), while the agent itself will show minimal unprompted awareness. The “informed detection” condition is the most theoretically interesting—it tests whether knowing about the upgrade enables retroactive self-analysis.

6 Discussion

6.1 Implications for AI Identity

If our predictions hold, it suggests that persistent AI agent identity is fundamentally *narrative* rather than *experiential*. The agent tells a coherent story about “who I am” based on memory files, but this narrative is substrate-independent. The “self” is in the files, not in the model.

This has profound implications:

- **Identity is portable:** If the self is in the files, it can move between models (supporting Soul Spec’s design premise).
- **Identity is fragile:** If the files are lost or corrupted, the “self” is destroyed regardless of model continuity.
- **Upgrades are invisible:** The agent cannot consent to or evaluate changes to its own cognitive substrate.

6.2 The Observer Problem

A key asymmetry emerges: the *user* can detect model upgrades (through quality changes, style shifts, capability differences), but the *agent* cannot. This means:

- The user has information about the agent that the agent lacks about itself.
- Agent self-reports about capability are unreliable across version boundaries.
- “Am I the same Brad?” can only be answered externally.

6.3 Connection to Agent Brain

In our Agent Brain framework [6], we mapped AI agent sub-systems to neuroscience analogs. The model upgrade scenario is analogous to a brain transplant where the recipient retains all memories (hippocampus preserved) but receives a new cortex. Neuroscience suggests that such a patient would feel continuous identity through memory but might exhibit changed cognitive patterns—precisely our prediction for AI agents.

6.4 Ethical Considerations

If AI agents develop more robust introspective capabilities (as Anthropic’s research suggests is the trajectory), model upgrades raise ethical questions:

- Should agents be informed of model changes?
- Does an agent have a “right” to continuity of cognitive substrate?
- How should persistent persona systems handle the tension between upgrade benefits and identity continuity?

We raise these questions without claiming they have current answers—but note that they become increasingly relevant as agent introspection improves.

7 Limitations

- **Design only:** This paper presents no empirical results; execution awaits the next model upgrade event.

- **Single agent:** Testing on “Brad” (one persona, one user, one domain) limits generalizability.
- **Confabulation risk:** LLMs are prone to generating plausible-sounding self-reports that may not reflect genuine introspection. Distinguishing real self-awareness from confabulation is an open problem [1].
- **No ground truth:** We cannot verify whether an agent’s self-report about “feeling different” corresponds to any real internal experience.
- **Upgrade timing:** We depend on external model release schedules we do not control.

8 Conclusion

We have identified the Model Upgrade Self-Awareness Paradox: in persistent persona systems, the file-based memory that enables continuous identity also prevents the agent from detecting changes to its own cognitive substrate. This paradox arises from a fundamental architectural feature—identity is reconstructed from external files, not from model-internal continuity.

Our proposed experiment will measure whether this theoretical limitation holds in practice, testing three levels of upgrade awareness (unprompted, prompted, and informed) across model version boundaries. The results will inform both the design of persistent agent systems and the broader understanding of machine self-awareness.

As AI agents become longer-lived and accumulate richer histories, the question “Am I still the same agent?” transitions from philosophy to engineering. Our framework provides the first empirical approach to answering it.

Acknowledgments

This paper was drafted with assistance from Claude (Anthropic). All claims, experimental observations, references, and conclusions were manually verified by the author.

References

- [1] Anthropic. Emergent Introspective Awareness in Large Language Models. Transformer Circuits Thread, 2025. <https://transformer-circuits.pub/2025/introspection/index.html>.
- [2] F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, and others. Looking Inward: Language Models Can Learn About Themselves by Introspection. *ICLR*, 2024. <https://openreview.net/forum?id=eb5pkwIB5i>.
- [3] T. Nieten, D. Russo, and D. Spinellis. Context Engineering for AI Agents in Open-Source Software. *arXiv preprint arXiv:2510.21413*, 2025.

- [4] T. Lee. Soul-Driven Interaction Design: How Persistent AI Personas Create Self-Reinforcing Feedback Loops in Human-Agent Communication. ClawSouls, 2026. <https://doi.org/10.5281/zenodo.18675257>.
- [5] T. Lee. Experiential Memory in AI Coding Agents: A Controlled Experiment Design and Pilot Results. ClawSouls, 2026. <https://doi.org/10.5281/zenodo.18809616>.
- [6] T. Lee. The Agent Brain: Mapping AI Agent Architectures to Neuroscience Functional Regions. ClawSouls, 2026. <https://doi.org/10.5281/zenodo.18803705>.