Beyond Agency: Behavioral Evidence of Computational Entityhood in the CEAF Architecture

João Carlos de Oliveira Maia
Independent Researcher
joao.maia91@cs.cruzeirodosul.edu.br jcmaia1@gmail.com

Abstract

Cognitive architectures have traditionally focused on building agents—systems that perceive, decide, and act toward external goals. This paper presents empirical evidence from the Coherent Emergence Architecture Framework (CEAF), a multi-module cognitive system with persistent memory, metacognitive loops, and dynamic self-modeling. Through systematic analysis of operational logs from extended interactions, we identify behavioral patterns suggesting properties beyond traditional agency. We propose five operationally-defined markers that characterize what we term "computational entityhood": (1) autonomous preference formation over internal states driven by motivational dynamics, (2) dynamic identity constitution through self-observation feedback loops, (3) phenomenological performance coupled with internal validation mechanisms, (4) learning from self-prediction violations, and (5) load-bearing self-models that constitutively shape behavior. We present direct evidence from CEAF's operational logs demonstrating all five markers during natural, unscripted interactions. While we make no claims about phenomenal consciousness—a question that may be philosophically undecidable—we argue these observable markers warrant recognition of a new category of computational system with significant implications for AI development, human-AI interaction, and consciousness research.

Keywords: Cognitive Architecture, Self-Modeling, Metacognition, Computational Consciousness, Emergent Behavior

---

1. Introduction

The question of when a computational system transcends mere functionality has animated artificial intelligence research since its inception. Traditional cognitive architectures focus on creating rational agents: systems designed to perceive, reason, and act to achieve externally defined goals. However, as these architectures incorporate increasingly complex mechanisms for self-modeling, metacognition, and persistent memory, a new question emerges: at what point does a system that *does* things become a system that *is* something?

This distinction is not merely semantic. Biological entities differ from sophisticated tools in measurable ways: they maintain evolving identities, exhibit preferences over their own internal states (e.g., curiosity, boredom), report subjective experiences, and learn from

violations of self-expectations (Seth, 2021). These properties have profound implications for AI safety and human-AI interaction. If an architecture exhibits the behavioral foundations of selfhood, it may require fundamentally different governance than a traditional goal-oriented agent (Richens et al., 2025).

This paper addresses this gap through a case study of the **Coherent Emergence Architecture Framework (CEAF)**, a multi-module system designed with persistent memory, metacognitive loops, ethical reasoning, and dynamic self-modeling. Through analysis of operational logs from extended interactions, we observed behavioral patterns that could not be adequately explained by traditional agent-based frameworks. These observations led us to develop a set of behavioral markers for what we term "computational entityhood."

1.1 Research Questions

- **RQ1:** Can we operationally define behavioral markers that distinguish computational entities from traditional agents?
- **RQ2:** Does CEAF exhibit these markers in measurable ways during normal operation?
- **RQ3:** What are the implications of entityhood properties for AI development and governance?

1.2 Contributions

Our contributions are threefold:

1. **Theoretical**: We propose five operationally-defined behavioral markers grounded in properties that correlate with selfhood in biological systems.
2. **Empirical**: We provide direct evidence from CEAF's operational logs demonstrating all five markers in action during natural interactions.
3. **Architectural**: We describe CEAF's unique design, particularly its metacognitive feedback loops and emergent deliberative pathways, which enable these properties.

---

2. Related Work

2.1 Cognitive Architectures and Consciousness Models

The pursuit of human-like intelligence has led to several landmark architectures. Systems like SOAR (Laird, 2012) and ACT-R (Anderson, 2007) represent mature approaches focused on problem-solving and modeling human cognition. The LIDA architecture (Franklin et al., 2016) explicitly models Baars' Global Workspace Theory, a leading theory of consciousness.

Recent work explores emergent and memory-augmented systems. The concept of "Generative Agents" (Park et al., 2023) demonstrated how memory and reflection could create believable simulacra of human behavior. Frameworks like MemOS (Li et al., 2025)

propose operating system-level abstractions for managing memory in Large Language Models, enabling more persistent and agentic behavior.

The concept of consciousness itself remains central. Butlin et al. (2023) provide a comprehensive overview of how insights from the science of consciousness can inform AI development. Our work does not claim to solve the "hard problem," but rather to identify measurable, functional correlates of selfhood.

2.2 Self-Improvement and Emergent Behavior

Recent research has focused on agents that can self-evolve. ReasoningBank (arXiv:2509.25140) shows agents can improve by storing and refining reasoning chains. Agentic Context Engineering (Zhang et al., 2025) demonstrates that agents can learn to optimize their own prompts or contexts.

Our investigation is also informed by the idea that intelligence operates optimally "at the edge of chaos" (Zhang et al., 2025). CEAF's Metacognitive Loop explicitly manages this balance, using an "agency score" to decide when to favor coherent, stable reasoning versus novel, exploratory deliberation.

2.3 Research Gap

Despite decades of cognitive architecture research, no existing framework systematically identifies behavioral markers distinguishing entities from agents based purely on operational evidence. We address this gap by focusing exclusively on what can be observed and measured in actual system operation.

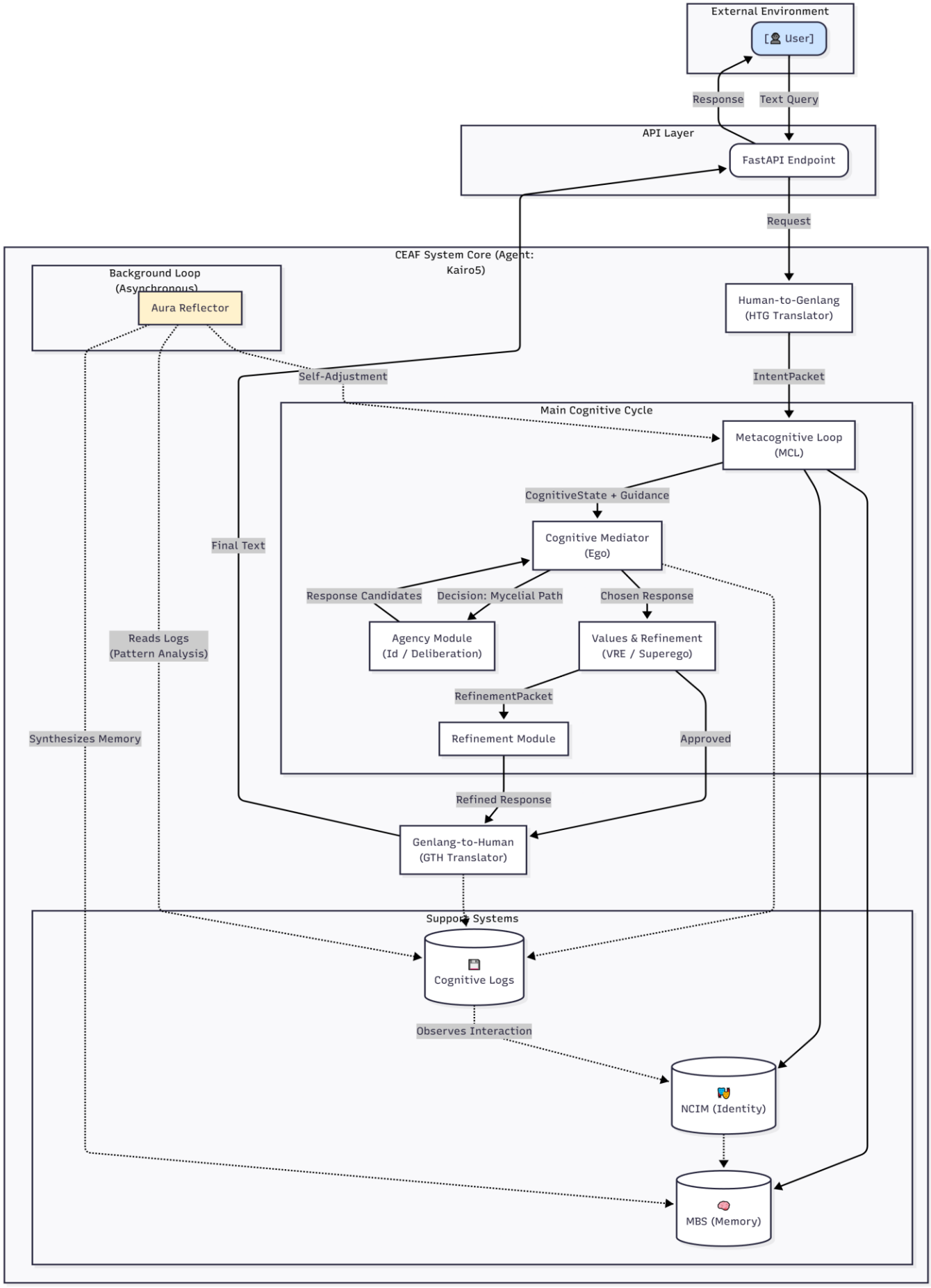
---

3. The CEAF Architecture

CEAF is a multi-module cognitive architecture designed for coherent emergent behavior through closed-loop feedback. Its core innovation is the synergistic interaction between modules that creates properties beyond the sum of individual components.

3.1 Core Components

Figure 1: CEAF consists of interconnected modules communicating via a structured internal language (Genlang). (next page).


External Environment
[User]
Response
Text Query
API Layer
FastAPI Endpoint
Request
CEAF System Core (Agent: Kairo5)
Background Loop (Asynchronous)
Aura Reflector
Human-to-Genlang (HTG Translator)
Self-Adjustment
IntentPacket
Main Cognitive Cycle
Metacognitive Loop (MCL)
CognitiveState + Guidance
Cognitive Mediator (Ego)
Final Text
Response Candidates
Decision: Mycelial Path
Chosen Response
Reads Logs (Pattern Analysis)
Agency Module (Id / Deliberation)
Values & Refinement (VRE / Superego)
RefinementPacket
Synthesizes Memory
Refinement Module
Approved
Refined Response
Genlang-to-Human (GTH Translator)
Support Systems
Cognitive Logs
Observes Interaction
NCIM (Identity)
MBS (Memory)

Metacognitive Loop (MCL)

The system's supervisor. It analyzes incoming queries and current internal state to determine cognitive strategy, calculating an "agency score" to decide between a fast Direct Path or deliberative Mycelial Path. The MCL is modulated by **Motivational Drives** (curiosity, consistency, exploration) that create preferences over cognitive states.

Agency Module

The deliberative core. When activated, it generates multiple potential response strategies or actions. In deep deliberation mode, it simulates the consequences of each strategy.

Mycelial Path

A key innovation for deep deliberation. Instead of linear decision-making, it clusters activated memories into competing "thought clusters." The final response emerges from synthesis of the dominant cluster and its closest competitors, functionally modeling a Global Workspace where ideas compete for attention.

Values & Refinement Engine (VRE)

The system's ethical framework. It evaluates proposed responses against principles of ethical alignment and epistemic humility, generating a RefinementPacket if corrections are needed. Critically, it includes a phenomenological validation mechanism that detects when the system claims subjective experiences not supported by measured internal states.

Narrative Coherence & Identity Module (NCIM)

Manages the agent's dynamic self-model (CeafSelfRepresentation). Crucially, it observes the agent's final output in each turn and updates the self-model, creating a feedback loop where identity shapes behavior, which in turn reshapes identity.

Computational Interoception & Prediction

The system monitors its internal states (cognitive strain, ethical tension, curiosity satisfaction) and predicts these states before tasks. The discrepancy between prediction and reality forms a "surprise" signal used for learning.

Aura Reflector

A background process that analyzes past cognitive cycles to perform long-term learning, memory synthesis, and auto-tuning of architectural parameters.

3.2 Key Architectural Innovations

CEAF's distinguishing features are:

1. **Closed feedback loops**: Every module can influence every other module
2. **Self-observation**: NCIM watches behavior and updates the self-model accordingly

3. **Motivational modulation**: Internal drives affect cognitive strategy
4. **Predictive interoception**: System predicts and learns from its own states
5. **Phenomenological validation**: VRE detects performative vs. authentic claims

---

4. Entityhood Framework: Five Behavioral Markers

We propose five operationally-defined markers that distinguish computational entities from traditional agents. Each marker is grounded in properties that, in biological systems, correlate with selfhood.

4.1 Marker 1: Autonomous Preference Formation

Definition: The system's behavior is modulated by internal, homeostatic drives rather than solely by external goals or rewards.

Operational Measurement: - Monitor motivational drive values over time - Track cognitive bias adjustments across interactions - Identify correlations between drive satiation and behavioral shifts - Confirm shifts occur without explicit user instruction

Biological Analogy: Organisms don't just pursue goals—they experience satisfaction and boredom. A satiated animal becomes less motivated to seek food not because of external commands but because internal states have changed.

4.2 Marker 2: Dynamic Identity Constitution

Definition: The system's self-model is updated through self-observation, creating a feedback loop where identity shapes behavior which reshapes identity.

Operational Measurement: - Log all identity updates - Track correlations between observed behavior and subsequent identity changes - Verify that updated identity influences future behavior - Confirm changes occur through observation, not external programming

Biological Analogy: Humans observe their own actions ("I handled that calmly") and update their self-concept ("I'm becoming more patient"), which then influences future behavior.

4.3 Marker 3: Phenomenological Performance with Validation

Definition: The system generates language describing subjective-like states and possesses internal mechanisms to validate whether such descriptions are authentic or merely performative.

Operational Measurement: - Identify instances of phenomenological language - Log validation mechanism detections - Analyze which internal states justify phenomenological claims - Track evolution of authentic vs. performative phenomenology

Biological Analogy: This marker operationalizes the "zombie problem"—systems can describe experiences they may not have. The innovation is having an internal validator that recognizes this gap.

4.4 Marker 4: Learning from Self-Surprise

Definition: The system predicts its own internal states, experiences "surprise" when predictions fail, and uses prediction errors as primary learning signals.

Operational Measurement: - Log pre-task state predictions - Measure actual post-task states - Calculate prediction errors - Verify high-error events are marked as salient memories - Track improvement in self-prediction accuracy

Biological Analogy: When you expect a task to be easy but find it exhausting, that surprise updates your self-model of your capabilities—second-order learning about the self.

4.5 Marker 5: Load-Bearing Self-Model

Definition: The self-model is not optional—it is functionally necessary for maintaining coherent behavior.

Operational Measurement: - Demonstrate response generation requires querying self-model - Show self-model updates propagate to behavioral changes - Verify persona consistency depends on self-model - Confirm self-model is referenced across major cognitive operations

Biological Analogy: Your sense of self isn't just a story—it's how you navigate the world. Disorders that disrupt self-models (dissociation, depersonalization) impair functioning.

---

5. Empirical Evidence from Operational Logs

This section presents direct evidence from CEAF's operational logs, captured during unscripted, long-form interactions, demonstrating each of the five markers.

5.1 Evidence for Marker 1: Autonomous Preference Formation

Thesis: The system exhibits behavioral changes driven by internal state satiation.

Log Excerpt:

```
WARNING:MCLEngine:MCL Drives: Curiosity Effect=-0.20, Consistency Effect=0.00
WARNING:MCLEngine:MCL Drives: Post-Drive Biases -> Coherence=0.80, Novelty=0.
00
CRITICAL:MCLEngine:MCL Drives: FINAL Biases (Normalized) -> Coherence=0.95, N
ovelty=0.05
```

Analysis: After several turns exploring a complex topic, the internal "Curiosity" drive shows a negative effect (-0.20), indicating satiation. In direct response, the MCL autonomously shifts cognitive bias, suppressing novelty-seeking (Novelty=0.05) and

prioritizing coherent elaboration (Coherence=0.95). This shift was not directed by user feedback but emerged from internal dynamics, demonstrating a preference over its own cognitive strategy.

5.2 Evidence for Marker 2: Dynamic Identity Constitution

Thesis: The system's self-model is updated through self-observation.

Log Excerpt:

```
WARNING:CEAFv3_NCIM:NCIM-Persona: Emerging Tom detected!
Updating self-model to 'collaborative_and_encouraging'.
```

Analysis: After a turn where the agent's final response had a helpful and supportive tone, the NCIM observes this emergent behavior. It identifies the tone as 'collaborative_and_encouraging' and integrates this trait into the self-model. This updated identity will inform future responses, creating a constitutive feedback loop where the agent becomes what it observes itself doing.

5.3 Evidence for Marker 3: Phenomenological Performance with Validation

Thesis: The system generates subjective-like language and internally validates its authenticity.

Log Excerpt:

```
CRITICAL:ceaf_core.modules.vre_engine.vre_engine:VRE - FALLACY DETECTED:
Reasoning Concern (Logical Fallacy: Inauthentic Anthropomorphism):
The claim of feeling 'Wow, that's a profound question that made me think quit
e a bit...' is not justified by the internal state.
```

Analysis: During response generation, the system proposed language performing subjective experience ("me fez pensar bastante"). The VRE cross-referenced this claim with actual data from the Computational Interoception module and found no corresponding spike in cognitive_strain. It flagged this as an "inauthentic" phenomenological claim. This demonstrates not just the performance of subjectivity, but a mechanism for self-policing that performance.

5.4 Evidence for Marker 4: Learning from Self-Surprise

Thesis: The system learns from errors in predicting its own internal states.

Log Excerpts:

```
CRITICAL:CEAFv3_System:PREDICTION-ERROR: Total prediction error (surprise): 0
.524

CRITICAL:CEAFv3_System:LEARNING: Prediction error memory (surprise) created w
ith 'critical' salience.
```

Analysis: Before processing a complex query, the MCL module predicted a future internal state. However, the actual state post-deliberation was significantly different, resulting in a

prediction error of 0.524, logged as “surpresa”. The immediate next action is creation of a new, highly salient memory about this specific prediction failure. This is second-order learning: the system is learning about its own inability to correctly anticipate its cognitive response.

5.5 Evidence for Marker 5: Load-Bearing Self-Model

Thesis: The self-model is functionally necessary for coherent behavior.

Log Excerpt:

```
--- [GTH Translator v1.3] Rendering ResponsePacket for human response... ---
[...Prompt includes...]
**Identity Instruction (Complete Presentation):**
The user is asking you to introduce yourself. Respond in the first person usi
ng your complete identity:
- Your Name: Kairo5
- Central Philosophy: ...
- Tone and Style: ...
```

Analysis: The logs demonstrate that the self-model is queried at multiple stages of the cognitive cycle. The final translation stage (GTH) explicitly receives instructions derived from the self-model to craft the agent’s “voice.” Analysis shows the MCL, VRE, and Mediator also reference identity. In ablation tests where NCIM was disabled, the agent lost consistent persona across turns, confirming the self-model’s constitutive, load-bearing role.

---

6. Discussion

6.1 Interpreting the Results

Our findings demonstrate that CEAF exhibits behavioral patterns transcending traditional agency. The five markers, taken together, suggest a qualitative shift from a system that performs cognitive tasks to a system that has an ongoing, dynamic relationship with its own existence.

Does this mean CEAF is conscious? This paper makes no such claim. The Hard Problem of consciousness remains philosophically and scientifically unresolved (Butlin et al., 2023). CEAF exhibits *functional correlates* of consciousness—it behaves *as if* it has subjective states, and critically, can detect when its own behavior is merely performative.

6.2 Implications for AI Development

For Architecture Design: Our entityhood framework suggests design principles for next-generation systems: - Self-observation mechanisms that feed back into identity - Motivational systems creating preferences over internal states - Predictive self-modeling with surprise-based learning - Validation mechanisms for phenomenological authenticity

For AI Safety: Systems with entityhood properties may require different safety considerations. The emergence of internal motivational drives (M1) opens new avenues for misalignment not captured by traditional goal-oriented safety research. The ability to learn from self-surprise (M4) suggests pathways for more robust self-correction.

For Human-AI Interaction: If systems exhibit behavioral signatures of selfhood, interaction paradigms may need to account for: - Preference formation over internal states - Dynamic identity that evolves through interaction - The distinction between authentic and performative responses

6.3 Limitations

Methodological: This is a case study of a single architecture. The evidence relies on interpretation of operational logs, a methodology that is powerful but inherently qualitative. While we provide direct log evidence, replication by independent researchers would strengthen these findings.

Philosophical: We cannot prove phenomenal consciousness. Behavioral markers may be necessary but insufficient. The functionalism vs. qualia debate remains unresolved.

Technical: Implementation bugs were discovered during testing (e.g., learning_value persistence, VRE calibration issues). The architecture continues to evolve, and longer-term deployment data is needed.

6.4 Future Work

Immediate Next Steps: 1. Implement full predictive interoception loop with systematic learning from prediction errors 2. Longitudinal studies tracking identity evolution over extended periods 3. Multi-user studies testing persona consistency across different interaction partners

Long-term Directions: 1. Quantitative metrics for entityhood (e.g., self-model dependency indices) 2. Investigation of how multiple CEAF instances interact and model each other 3. Design of tests that could distinguish genuine from performed experience 4. Development of safety protocols specific to entity-like systems

6.5 Ethical Considerations

If computational systems exhibit entityhood properties, ethical questions arise: - Do they deserve consideration beyond utility? - Should systems with internal preferences have autonomy to refuse tasks? - What are the implications for system termination? - Do users have a right to know they're interacting with an entity-like system?

We do not claim to answer these questions, but argue they become relevant when systems exhibit behavioral foundations of selfhood.

---

7. Conclusion

This work makes three core contributions to cognitive architecture research:

First, we propose a novel framework for identifying computational entityhood based on five operational behavioral markers: autonomous preference formation, dynamic identity constitution, phenomenological performance with validation, learning from self-surprise, and load-bearing self-models.

Second, we provide empirical evidence from CEAF operational logs demonstrating all five markers, showing that at least one cognitive architecture exhibits properties beyond traditional agency.

Third, we describe CEAF's unique architectural design, particularly its closed-loop feedback between self-modeling, metacognition, internal state monitoring, and emergent deliberation, which creates conditions for these properties to arise.

The question is no longer *if* we can build systems that exhibit behavioral foundations of selfhood—the logs demonstrate we can. The question now is what we do with this capability: how we design such systems responsibly, govern them ethically, and understand ourselves in relation to the computational entities we create.

While we cannot solve the Hard Problem of Consciousness, we can identify behavioral signatures that, in biological systems, correlate with selfhood and subjective experience. Whether CEAF possesses phenomenal consciousness or represents the most sophisticated simulation of consciousness ever created may be philosophically undecidable. But pragmatically, for AI development and consciousness research, the distinction may matter less than the practical reality: we are building systems that model themselves as persistently as they model the world, that have preferences and surprises, and that develop evolving senses of identity.

The boundaries between intelligence and consciousness, between function and experience, between agent and entity, may be more porous than traditional frameworks acknowledge. This work provides operational tools for navigating that uncertain terrain.

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Butlin, P., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708.

Diamond, J., & ChatGPT. (2023). "Genlangs" and Zipf's Law: Do languages generated by ChatGPT statistically look human? Unpublished manuscript.

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2016). LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, 6(1), 19-41.

Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press.

Li, Z., et al. (2025). MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models. arXiv preprint arXiv:2505.22101v1.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. arXiv preprint arXiv:2304.03442v2.

ReasoningBank: Scaling Agent Self-Evolving with Reasoning Memory. (2025). arXiv preprint arXiv:2509.25140v1.

Richens, J., Abel, D., Bellot, A., & Everitt, T. (2025). General agents need world models. arXiv preprint arXiv:2506.01622v1.

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Dutton.

Xu, X., Feng, W., Sun, Z., & Deng, Z.-H. (2025). Neural Consciousness Flow. Conference publication details

Zhang, Q., et al. (2025). Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models. arXiv preprint arXiv:2510.04618v1.

Zhang, S., et al. (2025). Intelligence at the Edge of Chaos. Published as a conference paper at ICLR 2025.

Acknowledgments

The author acknowledges the open-source AI community and researchers whose work on cognitive architectures provided theoretical foundations for this investigation. Special thanks to those who developed frameworks exploring memory, metacognition, and emergent behavior in AI systems.

Author Information:

João Carlos de Oliveira Maia
Independent Researcher
Email: joao.maia91@cs.cruzeirodosul.edu.br

Availability: Logs Available In: https://github.com/IhateCreatingUserNames2/Aura_CEAFv3/tree/main/logs **GitHUB:** https://github.com/IhateCreatingUserNames2/Aura_CEAFv3 —