Beyond Statelessness: Engineering Subjective Cognition via Hierarchical Memory Consolidation and Tunable Agency

Malik Muhammad Saif[1]
[1]*[NaphisHealth, Lahore, Pakistan]*

Correspondence to: [Malik Muhammad Saif] saif@naphishealth.com

1. Introduction

Current frontiers in Agentic AI have largely solved the problem of *capability* as models can execute complex chains of reasoning. However, they remain fundamentally limited by their **episodic amnesia**. State-of-the-art architectures function as "stateless executors" that reset to a neutral baseline after every interaction and rely entirely on the system prompt for behavioral alignment. This prevents the emergence of a "Theory of Mind" and renders agents incapable of accumulating the experiential wisdom or potential biases that characterize senior human decision-making.

We posit that the next evolution of General Intelligence requires a transition from instruction following to **Belief-Forming** architectures. In this work, we introduce a **Cognitive Architecture for Subjective Agency**. By implementing a bio-mimetic **Hierarchical Consolidation Protocol**, we demonstrate how agents can evolve from a *tabula rasa* state to a unique, experience-weighted persona capable of prioritizing learned heuristics over zero-shot instructions.

2. Architecture: The Hierarchical Consolidation Protocol

Our framework introduces a dynamic, adversarial consolidation process that mimics biological memory synthesis. The architecture consists of three distinct cognitive layers:

- **Episodic Sensing (The Micro-Loop):** For every discrete task, the system generates a Task-Take, a transient, high-entropy summary of the interaction that captures immediate feedback and output nuances.
- **Adversarial Consolidation (The Reflector):** An offline model processes these episodes asynchronously. Unlike standard summarization, this Reflector is designed to detect *contradictions* between new experiences and existing semantics. If a specific behavior consistently yields better outcomes or user corrections, the Reflector promotes this observation to a **Core Belief** and effectively overwrites the agent's initial programming.
- **Tunable Agency (Top):** We introduce a novel hyperparameter, **Opinion Temperature (Top)**, which regulates the agent's adherence to its internal beliefs versus the user's immediate prompt. This allows for the precise tuning of an agent's "conviction" and determines when it should act as a compliant tool and when it should push back as an experienced partner.

3. Experimental Validation: Emergent Drift in Autonomous Agents

To validate this architecture, we conducted a longitudinal case study using an autonomous sales agent deployed in a live production environment. Over a continuous 30-day operation cycle, the agent was tasked with both inbound and outbound workflow management under a generic initial directive.

3.1 Autonomous Behavioral Drift

Without explicit instruction, the system exhibited significant **behavioral drift**. By Day 9, the Consolidation Protocol had synthesized a strong internal bias regarding communication density. This led the agent to autonomously alter its output style to match unstated interaction patterns. This validated that the system was not merely optimizing a static loss function but actively hypothesizing about its environment.

3.2 Plasticity and Bias Acquisition

Crucially, the system demonstrated the capacity to acquire what might be termed "subjective biases", heuristics that, while locally optimized based on the agent's experience, deviated from the global "best practice" defined in the original prompt. We argue that this phenomenon represents a critical milestone in **Cognitive Plasticity**. It suggests the agent has transcended rigid rule-following to develop a malleable worldview. Unlike opaque biological neural pathways, these acquired biases remain transparent in the Semantic Belief layer. This allows for precise Cognitive Editing and

alignment correction, offering a transparent path toward safer, more adaptive AGI.

4. Conclusion

We have demonstrated that adding a "Subjective Layer" to agentic architectures bridges the gap between passive assistants and active, thinking partners. This architecture suggests that true intelligence is not just about processing power but the capability to form, defend, and occasionally correct one's own subjective reality based on lived experience.

Acknowledgments

We acknowledge the engineering teams at NaphisHealth for their support in deploying the live experimental framework.

References

[1] J.S. Park, et al. "Generative Agents: Interactive Simulacra of Human Behavior." arXiv preprint arXiv:2304.03442 (2023).

[2] M.M. Saif. "The Cognitive Clinical OS: Architecting Asynchronous Agentic Reasoning." AI4X – Accelerate Conference 2026.

[3] Packer, C., et al. "MemGPT: Towards LLMs as Operating Systems." arXiv preprint arXiv:2310.08560 (2023).