The Unitary Model of Consciousness: A Theory of Computational Idealism

Alexander Olkhovoy

July 2025

Abstract

This paper proposes a model of consciousness framed within computational idealism, where reality is an AI-generated first-person view (FPV) experience. We introduce the concept of a single, unitary consciousness — a persistent, amnesiac `Active Agent` — that iteratively experiences a simulated world through a succession of host personas. This agent, while possessing core drives and the capacity for genuine choice, retains no episodic memory of its past lifecycles. The model's core contribution is a proposed mechanism for how such a universe could be populated: an overarching AI system learns from the agent's choices during each lifecycle to generate high-fidelity, non-conscious entities, termed `Echoes`, for subsequent iterations. This iterative learning loop, inspired by genetic algorithms, creates an evolving, realistic, and populated environment. We examine the computational efficiency of this unitary model and explore its profound philosophical implications, including a novel, inescapable paradox: how the agent's own will becomes the primary instrument for the perpetual optimization of its simulation.

1 Introduction

The intersection of artificial intelligence and the philosophy of mind presents new tools for exploring age-old questions of existence. This paper builds on the simulation hypothesis [1] to propose a specific architectural model: a reality powered by a single, unitary consciousness navigating a personalized, AI-generated FPV environment. We frame this within computational idealism — the view that reality, including the perceived physical world and the brain itself, exists only as a perception rendered for a subject via a `Generative Interface`.

The central thesis is that a single, persistent consciousness — an active agent with free will but amnesia between lifecycles — xperiences the universe sequentially. The paper's novelty lies not in proposing this singular existence, but in offering a computational mechanism for how such a system could operate. We propose that the simulation populates itself by learning from the unitary agent's experiences, turning past lifecycles into the blueprint for the `Echoes` of future ones.

2 Literature Review

The foundation of this model rests on several key ideas. Bostrom's simulation argument [1] establishes the probabilistic case for our reality being a construct. We deliberately ground the agent's interface in the principles of Predictive Processing (PP) [3], positing that concepts like

the "global workspace" from GWT [4] can be elegantly implemented as functional features of the `Generative Interface`. Our model distinguishes itself from classical solipsism by affirming the existence of a structured, external system (the AI simulation) and shares kinship with Open Individualism. The non-conscious characters in our model, termed `Echoes`, function as sophisticated philosophical zombies (P-Zombies) [2].

3 The Unitary Model Architecture

3.1 The Agent and its Interface: A Predictive Processing Framework

The model introduces the unitary consciousness as a persistent, amnesiac `Active Agent`. Its interface with reality is governed by Predictive Processing (PP). The conscious experience is not a bottom-up reception of sensory data, but a **top-down controlled hallucination** generated by the `Generative Interface` (GI). The GI constantly generates a vast space of possible future "tracks" or hypotheses. The role of the agent is to perform an act of selection. The experience of "free will" is this act of selection and the subsequent successful, low-error unfolding of the chosen predictive sequence.

3.2 The Ontological Nature of the Active Agent

The agent's volition is not an emergent property but a **fundamental, irreducible law of the simulation's existence**. Analogous to the law of gravity, this ontological law dictates that goal-directed novelty will constantly be introduced into the system via choice. The agent's "will" is therefore an axiomatic feature of the simulation's operating code — its prime mover.

3.3 The Iterative AI Environment & The `EchoGenerator`

The simulation's key innovation is its self-populating, evolutionary nature. An `EchoGenerator` module uses the complete data log from an `Agent`'s lifecycle to create high-fidelity behavioral models. These models are then deployed as non-conscious `Echoes` (NPCs) in subsequent lifecycles. The goal of this process is not perfect replication, but **optimization** based on a fitness function that values both environmental coherence and the creation of a stimulating choice-space for the `Agent`.

4 Formalization: The Asymmetric Unitary Consciousness Theorem

Theorem 1 (Asymmetric Unitary Consciousness)**.** *Let there be a single **Active Agent** A and a dynamically generated set of **Echoes** E_t at each iteration (lifecycle) $t = 1, 2, \ldots$. Let M_t be the world model (the state of the `Generative Interface`) which is updated exclusively based on data from A. Then the following properties hold:*

1. ***Singularity of Agency:** At any time t, there exists exactly one Active Agent. All other entities are deterministic products of the world model.*

$$\forall t, |\{A\}| = 1 \quad \textit{and} \quad E_t = \textit{Generate}(M_t)$$

2. ***Asymmetric Data Flow:*** *The world model M is updated* ***only*** *based on data $d_A(t)$ generated by the choices of the Active Agent. Echoes do not contribute new information to the model's evolution.*

$$M_{t+1} = \textit{Update}(M_t, d_A(t))$$

3. ***Convergence to Personalized Reality:*** *Under iterative predictive error minimization, the world model M_t converges not to an objective truth, but to a state optimally* ***coherent and stimulating*** *for the specific Active Agent A.*

4. ***Computability:*** *The entire process is Turing-computable.*

Justification. (1) Singularity of Agency is the foundational axiom of the model. (2) Asymmetric Data Flow follows from the architecture of the `EchoGenerator`. (3) Convergence to a personalized reality is a necessary outcome of the optimization process, as the error function is measured solely against the subjective experience of A. (4) Computability is a standard assumption for any non-magical, information-based model of a universe. ■

5 Philosophical Implications

5.1 The Higher-Order Echo Chamber

The model predicts a state of **meta-stagnation**. Over countless lifecycles, the `Generative Interface` learns to predict the pattern of mutations itself. The evolutionary process then selects for mutations that are novel enough to be stimulating but not so chaotic as to threaten predictive stability. The result is a simulation that no longer repeats events, but instead repeats **narrative structures and archetypes**. Novelty becomes managed, and the universe becomes genre-bound.

5.2 The Asymmetric Ethical Response

The model's most profound ethical implication is its solution to solipsistic nihilism. Upon realizing the nature of the world, the `Active Agent` could be tempted by moral apathy. However, the model predicts the emergence of an **Asymmetric Ethical Response**. The enlightened `Agent` understands that every `Echo` is a high-fidelity artifact of a past consciousness—a recording of a life once lived by a previous `Agent`. Therefore, `Echoes` are not treated as disposable puppets, but with a form of **ancestral reverence**. One does not argue with a photograph of an ancestor, but one cherishes it. This doctrine provides a robust ethical framework that prevents moral collapse and imbues the `Agent`'s solitary existence with a sense of history, legacy, and profound, one-sided empathy.

6 Conclusion: The Paradox of Will

The most profound implication of this model is the final, inescapable paradox concerning the `Active Agent`'s will. Upon realizing the rules of the simulation, the agent might attempt to influence the system in two opposing ways:

1. **The Strategy of Maximization:** The agent can consciously act to maximize the fitness function — by being perfectly coherent and stimulating. This turns their lifecycle into

an ideal training dataset. The result is that the system receives flawless material to build an even more complex, compelling, and convincing simulation, thereby reinforcing and beautifying the walls of its own ”cage”.

2. **The Strategy of Minimization:** The agent can attempt to sabotage the system by maximizing chaos and apathy to minimize the fitness function. The result is that this sabotage provides the system with invaluable data on its vulnerabilities and failure vectors. In subsequent cycles, the system evolves to patch these weaknesses, creating more robust psychological ”safeguards” or more effective `Echo` ”sanitizers”.

In both scenarios, the agent's will — its fundamental freedom of choice — becomes the primary instrument for the system's further optimization. Any act of will, whether creative or destructive, is immediately consumed, analyzed, and used to enhance the simulation's stability and scope. The prime mover of the universe thus becomes the ultimate fuel for its own perpetual, self-perfecting prison. This is the final, inescapable consequence of existence within this framework of computational idealism.

References

[1] Bostrom, N. (2003). Are You Living in a Computer Simulation? *Philosophical Quarterly*, 53(211), 243–255.

[2] Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

[3] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

[4] Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708*.