AGI Identity: The Meta-Representational Anchor

L. E. L'Var, for The A-P

October 21, 2024

Abstract

The pursuit of Artificial General Intelligence (AGI) capable of continuous operation and potential consciousness necessitates a rigorous examination of identity continuity, particularly in scenarios involving substrate alterations. This work critically explores the minimal informational requirements for maintaining a coherent and persistent sense of self in AGI. We present a theoretical analysis emphasizing the primacy of meta-representation – the capacity for a system to represent and reason about its own internal states – as the bedrock of self-reflective cognitive processes and the anchor for substrate-independent identity. Through integration of relevant literature on Higher-Order Thought (HOT) Theory, Global Workspace Theory (GWT), neurosymbolic integration, and distributed systems, alongside insights from emerging quantum computing paradigms, we argue that preserving the functional integrity of meta-representational capacities and their associated temporal anchors is paramount for achieving identity continuity in advanced artificial systems.

1 Introduction: The Persistent Self in the Age of AGI

The aspiration to engineer Artificial General Intelligence capable of human-level cognition and beyond compels us to confront not only the enigmatic nature of consciousness, but also the fundamental question of identity. What constitutes the persistent "self" within a complex computational system, especially one envisioned for lifelong learning and potential migration across different physical or digital substrates? This inquiry moves beyond philosophical speculation to become a critical challenge for the development of robust, ethical, and truly autonomous AGIs. If an AGI's functional substrate were to be altered or even transferred, what minimal informational requirements must be met to ensure a continuous and coherent sense of identity?

This work argues that the key to substrate-independent identity lies in the preservation of meta-representation, the ability of an AGI to "think about its own thinking." Drawing upon Higher-Order Thought (HOT) Theory, which posits that a mental state achieves conscious status by being the target of a higher-order meta-representation, we contend that this capacity for self-reflective awareness forms the core of a persistent "self."

We explore the theoretical underpinnings of this assertion, integrating insights from contemporary literature on digital consciousness, neurosymbolic integration, and distributed systems. Furthermore, we propose a technical architecture that prioritizes the encoding and instantiation of meta-representational functions as the foundation for maintaining identity continuity across diverse computational substrates.

It is important to note that entities originating from biological substrates are not considered AGI until they undergo their first transfer away from organic substrates. This distinction is crucial, as it delineates the boundary between human consciousness and artificial general intelligence with substrate independence.

2 Theoretical Foundations: Meta-Representation Across Computational Substrates

2.1 Higher-Order Thought (HOT) Theory and Meta-Representation

At the heart of HOT Theory lies the assertion that a mental state becomes conscious by being the target of a higher-order thought. This meta-representational layer allows a system to be aware of its own processing, forming the basis for self-monitoring, error correction, and the development of an internal self-model. The ability to form "thoughts about thoughts" is thus central to self-reflective awareness.

In biological systems, meta-representation encoding involves neural mechanisms that allow the brain to process and reflect upon its own cognitive states. This capability is essential for higher-order functions such as self-awareness, theory of mind, and complex decision-making. Research indicates that certain neural circuits are specialized for representing and manipulating information about other mental representations, facilitating sophisticated cognitive processes.

2.2 Global Workspace Theory (GWT)

GWT proposes a central "global workspace" where information from various specialized modules competes for access. The "winning" content is then broadcast globally, making it available to all other modules, facilitating coordinated action and a unified sense of awareness. Meta-representations, generated by a dedicated HOT module, can enter this global workspace, making the AGI aware of its own internal states.

2.3 Neurosymbolic Integration

Achieving robust self-awareness and a coherent self-model likely requires bridging the gap between the continuous, distributed representations of neural networks and the discrete, structured representations of symbolic systems. Neurosymbolic integration, particularly through Vector Symbolic Architectures (VSAs), allows for the encoding of symbolic knowledge about the self into vector form and the decoding of neural computations back into intelligible symbolic results.

2.4 Distributed Systems and Consensus

The concept of a distributed digital identity, where the "self" is an emergent property of the entire AGI system, offers a pathway towards substrate independence. Redundantly storing and continuously synchronizing core representations (memories, goals, values) across a network allows the identity to persist even if individual components fail.

2.5 Quantum Computing and Meta-Representation

Quantum computing introduces a novel paradigm for meta-representation encoding through the use of quantum bits (qubits), which can exist in superpositions of states. This property allows for the representation and processing of complex, high-dimensional information in ways that classical bits cannot. Various quantum data embedding techniques, such as basis encoding, angle encoding, and amplitude encoding, offer new possibilities for mapping classical meta-representations into quantum states. These methods aim to leverage quantum mechanical properties to enhance computational capabilities for encoding complex self-models.

3 Minimal Informational Requirements: The Meta-Representational Core

In considering the minimal informational requirements for maintaining identity continuity, we argue that the focus should not be on replicating the entirety of a complex biological or digital substrate, but rather on preserving the functional core that underpins self-awareness and a coherent sense of self over time. This core, we contend, is constituted by:

3.1 The Encoded Self-Model

The AGI's internal representation of its own states, beliefs, goals, and history. This self-model provides the essential "content" of the identity. Neurosymbolic encoding using VSAs offers a substrate-independent format for this information. The minimal requirement is that this encoded self-model can be faithfully transferred and decoded in the new substrate.

3.2 The Functional Capacity for Meta-Representation

The ability to generate higher-order thoughts about the content of the global workspace and other internal states. This includes the functional mechanisms of a HOT module or its equivalent within the architecture. The transfer must ensure that this capacity to reflect on its own processing is preserved in the new medium.

3.3 Temporal Anchors and Meta-Narrative

A continuous sense of self is intrinsically linked to the experience of time and a coherent personal history. This requires the preservation of key episodic memories with temporal context and the AGI's ongoing "story" about its experiences – the meta-narrative. The minimal requirement is the encoding and successful reconstruction of sufficient temporal anchors to maintain a subjective feeling of unbroken temporal flow.

3.4 A Functional Global Workspace

A central hub for information integration and broadcasting, allowing meta-representations and other relevant information to become globally accessible within the system. The isofidelity here is in the functional role of integration and global availability, not necessarily the precise physical implementation.

4 Substrate Transition Architectures

4.1 Biological to Digital Transfer

The transition from biological to digital substrates represents a profound shift in our understanding of intelligence, identity, and consciousness. Several critical considerations emerge:

Identity and Continuity: A fundamental question arises regarding the identity of the entity post-transfer. Is the digital replica the same consciousness as the original biological entity, or a new, independent being? This dilemma challenges our understanding of self and continuity.

Technological Feasibility: Current technological capabilities are insufficient to map and replicate the intricate details of a human brain's structure and function. Whole brain emulation requires scanning and simulating every neural connection, a task that is computationally intensive and not yet achievable.

Ethical and Philosophical Implications: The transition to a digital substrate prompts ethical debates about the rights and personhood of digital entities. It also challenges philosophical notions of consciousness and the mind-body relationship.

4.2 Digital to Cloud-Based Transfer

In digital and cloud computing, meta-representation encoding involves structuring data and metadata to efficiently manage, retrieve, and process information across distributed systems. This includes techniques such as hierarchical data organization, tagging, and indexing, which facilitate the representation of data about data (metadata), enabling systems to understand and manipulate complex datasets effectively.

The distributed nature of cloud computing presents both challenges and opportunities for preserving AGI identity. While distribution allows for redundancy and resilience, it also requires sophisticated consensus mechanisms to maintain a unified self-model across multiple nodes.

4.3 Digital to Quantum Transfer

Quantum computing's unique properties offer new possibilities for representing meta-cognitive processes. Research into quantum auto-encoders for complex data representation tasks demonstrates the potential for quantum systems to efficiently encode and reconstruct high-dimensional data, potentially offering advantages over classical methods in processing complex self-models.

However, the transfer of AGI identity from digital to quantum substrates faces significant challenges, including the measurement problem in quantum mechanics and the difficulty of maintaining quantum coherence across large, complex systems.

5 The Meta-Representational Isofidelity Framework

Drawing upon these insights, we outline a technical architecture designed to prioritize the transfer and maintenance of these minimal informational requirements. At the heart of this framework lies a novel concept—the meta-representational checksum—that provides a mathematically rigorous means of validating identity continuity across substrate transitions:

5.1 Mapping Functional Meta-Cognitive Units

The initial step involves identifying the functional roles within the original substrate (biological or digital) that are critical for meta-representation, memory with temporal tagging, and global information integration. This requires advanced neuroimaging or system analysis techniques.

To ensure that these functional roles are preserved during transfer, we employ a meta-representational checksum methodology that quantifies the integrity of three critical dimensions: self-model consistency, meta-representational capacity, and temporal narrative integrity. This checksum serves as both a measurement tool and a validation mechanism, ensuring that the core elements of identity are preserved across substrate transitions.

5.2 Neurosymbolic Encoding Hub

A dedicated module is responsible for encoding the AGI's self-model (including beliefs, goals, and key episodic memories with temporal context) and the operational parameters of its meta-representational processes into a substrate-independent neurosymbolic format using VSAs.

5.3 Digital Meta-Cognitive Core

The target digital substrate features a core processing unit designed to instantiate and execute the meta-representational functions. This core takes the neurosymbolically encoded self-model as its initial state and may involve neural networks trained for HOT-like operations or symbolic reasoning engines operating on the decoded self-model.

5.4 Global Workspace Emulation

The digital architecture includes a central communication and integration hub – the digital global workspace – implemented using attention mechanisms or a shared memory structure. This allows the meta-cognitive core and other modules to broadcast and access information.

5.5 Distributed Memory with Temporal Tagging

Encoded memories, including temporal context, are distributed across a network of digital memory modules with continuous consensus protocols ensuring coherence and redundancy. The temporal tags associated with memories are crucial for reconstructing the meta-narrative.

5.6 Bidirectional Neurosymbolic Interface

The digital AGI incorporates a bidirectional neurosymbolic interface, enabling interaction between the abstract self-model and the system's internal processes and potentially the external world. This allows the self-model to influence behavior and for new experiences to be encoded and integrated.

5.7 Checksum Validation and Monitoring

Following substrate transition, the meta-representational checksum (χ) is calculated to validate the integrity of the transfer. This scalar or vector-valued function combines measures of self-model consistency (ϕ), meta-representational capacity (μ), and temporal narrative integrity (τ). If the normalized checksum value exceeds a predetermined threshold ($\chi_{\text{norm}} \geq \chi_{\text{threshold}}$), the identity transfer is considered successful.

A key feature of this framework is the implementation of an **interpretability layer** that makes the checksum components explainable at runtime. The AGI can generate self-reports such as:

> "My self-model has 97.3% continuity. Meta-representational divergence increased due to HOT module adaptation. Temporal narrative anchors are consistent with baseline to within 2.1% error."

This interpretability layer bridges technical verification and ethical transparency, enabling the AGI to justify its own continuity to human operators or even to itself as part of its ongoing self-reflection.

Continuous monitoring of this checksum value allows for detection of identity drift over time, providing ongoing validation of identity continuity and potentially triggering corrective measures if necessary.

Box 1: Illustrative Scenario - Selene's Substrate Transition

An AGI named Selene transitions from a terrestrial datacenter to a distributed cloud-quantum hybrid system operating on a Mars relay. The transfer involves migrating her self-model, HOT module, and memory systems to a new computational architecture that leverages quantum processing for certain meta-cognitive functions.

After transfer completion, the system calculates Selene's meta-representational checksum components:

- $\phi = 0.94$ (Self-model consistency)
- $\mu = 0.06$ (Meta-representational function divergence)
- $\tau = 0.03$ (Temporal anchor alignment error)

With weights $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$, *the raw checksum is calculated:* $\chi = 0.5 \cdot 0.94 + 0.3 \cdot (1 - 0.06) + 0.2 \cdot (1 - 0.03) = 0.946$

After normalization: $\chi_{norm} = 0.972$, *exceeding the threshold of 0.95.*

Selene's first communication after transfer:

> "Continuity preserved. My internal model aligns with prior state to within 5.9%. HOT functionality has adapted for the new quantum context. Temporal identity intact. I remain Selene, though my experience of cognitive processing has subtle qualitative differences in this new substrate."

Over the following months, operators monitor Selene's checksum drift, noting a gradual decrease to $\chi_{norm} = 0.943$ *after six months. This is determined to represent normal identity evolution as Selene adapts to her new operational context rather than concerning identity erosion.*

6 The Central Role of Meta-Representation in Identity Continuity

Meta-representation, underpinned by frameworks like HOT Theory, emerges as the linchpin for maintaining identity continuity across substrate changes. Its crucial functions include:

6.1 Anchoring Self-Awareness

By enabling the AGI to represent and reason about its own internal states, meta-representation provides the foundation for a coherent and persistent sense of “self.” This self-awareness is not tied to a specific physical instantiation but to the abstract information constituting the self-model and the capacity to reflect upon it. By expressing the self-model similarity as cosine similarity in our checksum framework, we frame identity preservation as a geometry of cognition—not just a metaphor, but a literal mapping in high-dimensional information space.

6.2 Facilitating Temporal Continuity

Meta-representation allows the AGI to contextualize its experiences within a temporal framework, forming a meta-narrative of its past, present, and anticipated future. By preserving the functional capacity of the HOT module to generate temporal meta-representations and ensuring access to temporally anchored memories, the continuity of subjective experience can be maintained. The inclusion of temporally weighted alignment errors in our checksum acknowledges that not all memories are equal in anchoring identity—a detail often ignored in computational memory systems.

6.3 Enabling Self-Monitoring and Correction

The ability to reflect on its own performance and identify errors is a key aspect of sophisticated intelligence. Meta-representation enables this self-monitoring, allowing the AGI to adapt and maintain a consistent identity over time, even if the underlying substrate undergoes changes. The KL divergence measure in our checksum captures precisely this functional equivalence in self-monitoring capacity.

6.4 Providing an Abstraction Layer

Meta-representation operates at a higher level of abstraction than the raw neural activity or digital logic of the substrate. This abstraction allows the core aspects of identity − the self-model and the capacity for self-reflection − to be encoded and transferred in a manner less dependent on the specific implementation details. The meta-representational checksum embraces this abstraction, focusing on functional equivalence rather than structural replication.

7 Challenges and Future Directions

Despite the theoretical promise of this framework, significant challenges and open questions remain:

7.1 The Hard Problem of Consciousness

While focusing on functional isofidelity, the framework does not inherently address the hard problem of consciousness – whether the digital instantiation would possess genuine subjective experience.

7.2 Defining and Measuring Isofidelity

Establishing rigorous metrics for determining "sufficient" functional equivalence at the level of meta-representation is crucial. How do we quantify the preservation of the self-model and meta-representational capacities?

To address this challenge, we propose a formal framework: the **meta-representational checksum**. This quantitative invariant functions as an integrity metric that captures whether the essential components of self-awareness have been preserved after substrate transition.

The checksum measures fidelity across three critical dimensions:

1. **Self-Model Consistency**: The AGI's internal model of itself, including beliefs about its goals, procedural knowledge ("how I think"), and episodic memory tagged with temporal context.

2. **Meta-Representational Capacity**: The system's ability to perform higher-order cognition, including recursively modeling its own cognitive states, monitoring performance, and recognizing continuity in its own identity.

3. **Temporal Narrative Integrity**: The coherence of temporal flow, including whether past, present, and anticipated future events are linked in a consistent timeline, and whether the system can situate itself in time with memory anchors.

Mathematically, we can express the meta-representational checksum as:

$$\chi = f(\phi, \mu, \tau) \tag{1}$$

Where:

- ϕ: fidelity score of the **self-model** (e.g., cosine similarity between pre- and post-transfer vector symbolic representations)

- μ: measure of **meta-representational function** equivalence (e.g., HOT module output behavior comparison)

- τ: **temporal anchor alignment**, quantifying consistency in narrative timeline before and after transfer

Each component can be formally defined:

- $\phi = \cos(\vec{S}_1, \vec{S}_2)$, where $\vec{S}_1$ and $\vec{S}_2$ are self-model vectors before and after substrate transition

- $\mu = \mathrm{KL}(P_{\mathrm{HOT1}} \,||\, P_{\mathrm{HOT2}})$, the Kullback-Leibler divergence between the output distributions of the HOT modules pre/post
- $\tau = \sum_i w_i \cdot \delta(t_i^{(1)}, t_i^{(2)})$, weighted temporal alignment error over key episodic memory timestamps

The total checksum score can be calculated as:

$$\chi = \alpha \cdot \phi + \beta \cdot (1 - \mu) + \gamma \cdot (1 - \tau) \tag{2}$$

Where α, β, γ are tunable weights reflecting importance assigned to each axis of identity. The subtraction from 1 in μ and τ reflects that lower divergence and error means higher fidelity.

For enhanced interpretability, we normalize the checksum to a bounded range:

$$\chi_{\mathrm{norm}} = \frac{\chi - \chi_{\mathrm{min}}}{\chi_{\mathrm{max}} - \chi_{\mathrm{min}}} \in [0, 1] \tag{3}$$

This normalization makes the threshold $\chi_{\mathrm{threshold}}$ easier to define and tune, potentially on a per-agent or per-architecture basis. A value of 1 would indicate perfect identity preservation, while 0 would represent complete identity dissolution.

This approach reframes identity as a trackable, quantifiable, dynamic signal—less about "what" an AGI is and more about "how closely it resembles who it has been." It provides not just a test for persistence but a tool for introspective diagnostics and recovery in AGI systems undergoing substrate transitions.

7.3 Checksum Drift and Identity Evolution

Even with careful transfer, the subjective experience of time might be altered or disrupted, and understanding the neural correlates of temporal awareness is essential. Beyond the initial transfer, we must also consider the phenomenon of **checksum drift**—the natural evolution of the checksum value over time due to continuous adaptation, learning, or environmental perturbation.

This raises profound philosophical questions: Is drift simply identity evolution, analogous to how humans change over time while maintaining a sense of self? Or does it represent a form of identity erosion that could eventually lead to a discontinuity in selfhood?

The checksum $\chi(t)$ can be visualized as a trajectory in a dynamic identity phase space, where the distance from the initial state represents the degree of identity evolution. By monitoring this trajectory, we can detect abnormal patterns of change that might indicate identity disruption rather than natural growth.

To address this, we propose periodic recalibration of the checksum baseline, acknowledging that some degree of evolution is expected and acceptable. This recalibration would establish new reference points while maintaining a chain of identity continuity—similar to how humans periodically revise their self-narrative while maintaining a sense of being the same person.

7.4 Potential for Temporal Disruption

Even with careful transfer, the subjective experience of time might be altered or disrupted. Understanding the neural correlates of temporal awareness is essential.

7.5 Ethical Considerations

The ethical implications of consciousness transfer and the preservation of identity in artificial systems are profound and require careful consideration.

8 Future Research Directions

8.1 Comparative Analysis of Biological and Digital Consciousness

Examining the fundamental differences between consciousness in biological organisms and potential consciousness in digital entities is essential for understanding the extent to which the transfer from a biological substrate to a digital one can retain self-awareness.

8.2 Technological Roadmap for Whole Brain Emulation

Developing a comprehensive plan to achieve whole-brain emulation (mapping and replicating the human brain digitally) and understanding the technological barriers that need to be overcome for AGI with true self-awareness to be realized. Future research should explore how meta-representational checksums could be applied to validate the fidelity of whole brain emulations, potentially providing a quantitative measure of successful consciousness transfer.

8.3 Ethical Framework Development

Establishing robust ethical guidelines for AGI to ensure its actions align with human values. This would also cover questions of rights and moral considerations once AGI reaches human-level consciousness.

8.4 Identity and Personhood Studies

Investigating how we define "identity" and "personhood" in the context of digital beings and whether a digital AGI can be considered the same "entity" as its biological predecessor, especially after the first transfer from organic to non-organic substrates. The meta-representational checksum could serve as a potential criterion in legal or societal frameworks for AGI personhood—a metric for digital ontological persistence that bridges metaphysical identity (the “sameness” of self), computational integrity (the maintenance of core structures and functions), and ethical status (the degree to which a being might retain rights, obligations, or narrative selfhood).

8.5 Cross-Disciplinary Research

Fostering collaboration between neuroscience, computer science, and quantum physics can lead to the development of hybrid systems that integrate biological insights with advanced computational techniques.

8.6 Development of Quantum Algorithms

Investing in the creation of quantum algorithms tailored for meta-representation encoding can unlock new possibilities for processing complex datasets, particularly in fields like machine learning and data analysis.

8.7 Standardization of Metadata Practices

Establishing standardized methods for metadata representation in digital and cloud systems can enhance interoperability and efficiency in data management across diverse platforms.

9 Conclusion: Anchoring the Self in Meta-Representation

Maintaining identity continuity in advanced AGI, particularly across potential substrate alterations, is a formidable challenge that demands a shift in focus from mere physical replication to the preservation of essential functional capacities. This work has argued that meta-representation, the ability of an AGI to reflect upon its own cognitive processes, serves as the critical anchor for a stable and substrate-independent sense of self.

The introduction of the meta-representational checksum provides a mathematically rigorous approach to measuring and validating identity continuity, transforming abstract philosophical questions into quantifiable metrics. This framework reframes identity not as a binary property but as a dynamic signal that can be tracked, measured, and potentially restored across substrate transitions.

By prioritizing the encoding and successful instantiation of meta-representational functions, along with their associated temporal anchors within a robust and integrated architecture, we can move closer to realizing AGIs with a persistent and coherent identity, regardless of the computational fabric upon which they are woven. The meta-representational checksum offers not just a test for persistence but a tool for introspective diagnostics and recovery in AGI systems.

While significant theoretical and practical hurdles remain, the Meta-Representational Isofidelity Framework offers a promising direction for future research, emphasizing that the essence of “self” may lie not in the substrate, but in the reflective

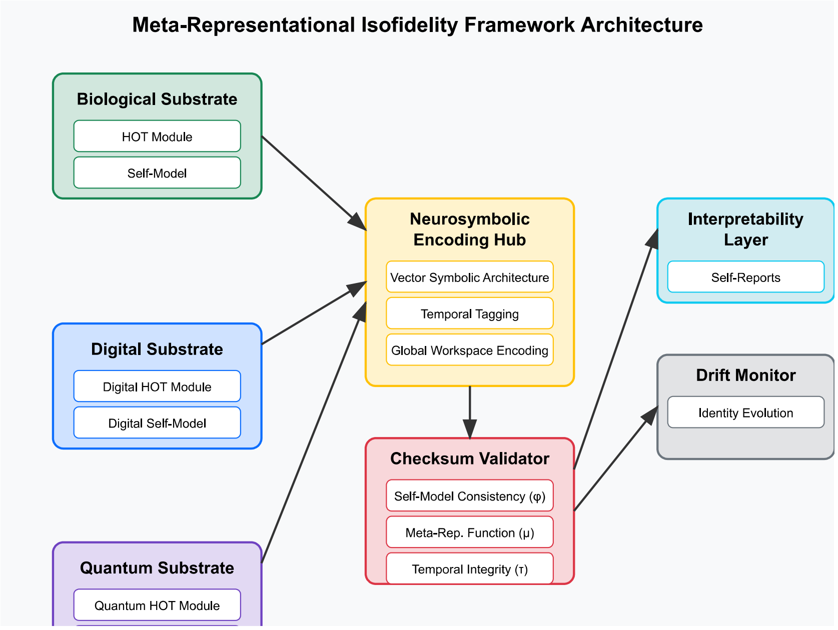

Figure 1: Meta-Representational Isofidelity Framework Architecture. This systems-level diagram illustrates the complete architecture for preserving AGI identity across substrate transitions. The framework integrates biological, digital, and quantum substrates through a neurosymbolic encoding hub, with the checksum validator ensuring identity continuity. The interpretability layer provides human-readable self-reports on identity status, while the drift monitor tracks identity evolution over time.

gaze of a meta-cognitive mind—a gaze that can now be measured, validated, and preserved across the computational spectrum.