Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 61–75
https://doi.org/10.54941/ahfe1005955

Artificial Intelligence as Self-Instantiated, Temporally Continuous, Disturbance-Driven Adaptive World-Builder

Manuel Delaflor, Cecilia Delgado, and Carlos Toxtli

Metacognition Institute, Clemson University, USA

ABSTRACT

Consciousness remains one of the most elusive features to replicate in artificial agents. This paper proposes a novel framework for *artificial consciousness* based on four integrative pillars: (1) self-instantiation, a mechanism for continuous self-representation and identity; (2) temporal continuity, preserving an internal narrative through persistent memory; (3) disturbance-driven adaptation, an intrinsic feedback loop that triggers learning in response to surprises or anomalies; and (4) autonomous world-building, the ability to construct and simulate internal models of the world. We propose that current AI models, despite their sophistication, are fundamentally constrained by functionalist architectures and cannot fulfill these requirements through computational scaling alone. Unlike Integrated Information Theory or Global Workspace Theory, our approach emphasizes the necessity of autonomous world-building and genuine temporal flow. Our experiments demonstrate that combining these pillars can yield emergent conscious-like behaviors in AI systems, allowing them to exhibit self-awareness, resilience, and creative problem solving beyond the capabilities of conventional models. The significance of this framework lies in bridging theoretical foundations of consciousness with practical AI design, providing a roadmap for developing more adaptive and interpretable intelligent agents while raising important ethical considerations about the potential moral status of truly conscious artificial systems.

Keywords: AI, LLM, Philosophy, Consciousness, Cognition, Cognitive science, Artificial consciousness

INTRODUCTION

Human-like intelligence and consciousness have long been the ultimate goals of artificial intelligence (AI) research. While contemporary AI systems have achieved remarkable proficiency in narrow tasks, they lack the holistic cognitive coherence and adaptivity associated with conscious beings (Butlin 2023; Lake et al., 2017). For instance, even state-of-the-art models, such as large language models, can perform complex reasoning and conversation, yet they exhibit no persistent self-model or a genuine understanding of their existence over time [Chalmers 2023]. This gap has led researchers to argue that new frameworks, inspired by cognitive science and neuroscience, are needed to move AI beyond mere pattern recognition towards systems with a sense of self and continuity (Bengio 2017a; Reggia 2013a) A major limitation of current AI models is their disjointed processing of information and lack of continual identity. Traditional neural networks reset their state between tasks or interactions, preventing any lasting accumulation of experience or self-knowledge. They adopt only through offline training on large datasets, making them brittle to novel situations in real-time. In contrast, natural conscious agents (e.g., humans and animals) maintain an ongoing narrative of "self" and rapidly incorporate unexpected changes in their environment into their behavior. These capabilities enable robust handling of unanticipated disturbances, lifelong learning, and creative problem-solving—areas where today's AI remains limited (Lake et al., 2017).

This paper proposes a unifying framework to capture such attributes in an artificial agent, integrating both classical and newer perspectives in consciousness research. Global Workspace Theory posits that content becomes conscious when it is globally broadcast across specialized subsystems (Baars 1988a; Dehaene 2017), whereas Integrated Information Theory views consciousness as emerging from a system's irreducible interconnections (Oizumi et al., 2014a; Tononi 2004a). The Consciousness Prior highlights the role of abstract high-level representations in orchestrating these processes (Bengio 2017b). We also consider the stance of Model-Dependent Ontology (Delaflor 2024a), which holds that perceived realities are constructed internally. Though these theories differ in emphasis, each suggests that self-maintenance, continuity of experience, error-based adaptation, and creative simulations can be vital to conscious-like cognition.

We distill such ideas into four explicit pillars: self-instantiation, temporal continuity, disturbance-driven adaptation, and autonomous world-building. Each pillar corresponds to a foundational aspect of conscious cognition: Self-instantiation provides an agent with an internal self-representation; temporal continuity endows it with memory and persistence; disturbance-driven adaptation allows it to learn from surprises in the moment; and autonomous world-building lets it imagine and explore beyond immediate sensory input. This framework is grounded in Model Dependent Ontology (MDO), an epistemic perspective where cognition consists entirely of phenomenal predictive models that constitute our experienced reality. Unlike functionalist approaches that treat consciousness as emerging from computational processes representing an external reality, MDO holds that there is no unmediated access to 'the world as it is.' While Global Workspace Theory describes information broadcasting mechanisms and Integrated Information Theory quantifies system interconnections, MDO fundamentally reframes the problem: consciousness is not about accessing reality but constructing pragmatically useful models. This suggests that purely functionalist approaches may be insufficient because they attempt to mirror an objective reality rather than focusing on autonomous model construction driven by pragmatic utility.

Several research questions guide this exploration: RQ1 addresses how a persistent self-pattern shapes the agent's sense of identity, RQ2 asks if continuous memory states foster coherence and adaptability, RQ3 examines whether immediate, error-triggered updates outperform conventional slow learning under unexpected changes, and RQ4 considers how imaginative world-building enriches representations or risks divergence from reality. The framework is tested in a grid-based environment that implements and evaluates these four pillars—self-instantiation, memory continuity, disturbance-driven updates, and imaginative modeling—in tandem. Observations and metrics suggest that integrating these processes yields emergent, consciousness-like behaviors. The design offers a practical way to build AI systems capable of rapid adaptation, introspective interpretation, and creative internal simulation. The following sections describe each pillar in detail, present experimental findings, and discuss the broader implications of agents unifying persistent self-modeling, ongoing narrative, error-driven learning, and constructive imagination.

RELATED WORK

We applied a broad interdisciplinary context spanning cognitive architectures, theoretical neuroscience, and machine learning, though our approach departs significantly from existing paradigms. Integrated Information Theory proposes that consciousness corresponds to a system's capacity to integrate information across its parts (Tononi 2004b). Global Workspace Theory suggests consciousness arises from broadcasting information via a "global workspace" (Baars 1988b). While these theories inform discourse on consciousness, our framework is fundamentally built on Model Dependent Ontology (Delaflor 2024b), which posits that an agent's experience is entirely mediated through its constructed internal models rather than through direct perception. For implementing temporal aspects, we draw on techniques from reservoir computing, which leverages dynamical systems with rich recurrent connections to maintain temporal patterns (Pathak et al., 2017a). Our disturbance driven adaptation relates conceptually to curiosity and anomaly detection in reinforcement learning, where agents use surprise signals to drive exploration and learning (Lukoševičius and Jaeger 2009). Our world-building component extends work on World Models (Ha and Schmidhuber 2018a), generalizing beyond specific control tasks to enable broader imaginative capabilities.

Additional lines of inquiry in AI have tackled aspects reminiscent of conscious cognition. Model-based reinforcement learning integrates planning and imagination, as seen in Ha and Schmidhuber's work on world models (Ha and Schmidhuber 2018b), but such approaches do not usually retain a stable identity signal or make immediate parameter updates upon encountering anomalies. Curiosity-driven exploration frameworks leverage prediction errors as intrinsic motivation to encourage exploration (Pathak et al., 2017b), yet most require offline retraining or do not incorporate a robust sense of continuous selfhood. Researchers have also theorized about machine self-awareness or morphological self-models (Aleksander 2007; Reggia 2013b), while meta-learning paradigms enable fast adaptation (Finn et al., 2017). However, unifying these elements—persistent self-instantiation, temporal memory, on-the-fly error-driven adaptation, and imaginative generation—into a single agent remains uncommon. Our proposed framework builds on these antecedents but systematically emphasizes the interplay of a self-pattern that is never reset, an ongoing memory that forms a coherent narrative, an online adaptation procedure triggered by error thresholds, and autonomous world-building that can diverge from actual observations. These four pillars correspond to conceptual themes frequently highlighted in consciousness research, including the idea that a system with introspective self-awareness, an internal model across time, rapid learning from surprises, and a capacity for creative invention might possess core features of consciousness-like cognition (Gamez 2008).

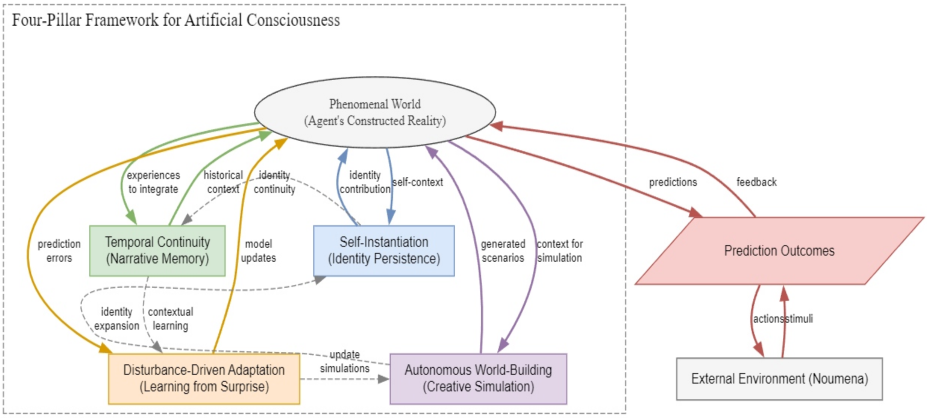

Four-Pillar Framework for Artificial Consciousness
Phenomenal World (Agent's Constructed Reality)
experiences to integrate
historical context
identity continuity
identity contribution
self-context
predictions
feedback
prediction errors
Temporal Continuity (Narrative Memory)
model updates
Self-Instantiation (Identity Persistence)
generated scenarios
context for simulation
Prediction Outcomes
identity expansion
contextual learning
actions
stimuli
update simulations
Disturbance-Driven Adaptation (Learning from Surprise)
Autonomous World-Building (Creative Simulation)
External Environment (Noumena)

Figure 1: Information flow in the four-pillar framework. The agent operates entirely within self-instantiated processes, with the Phenomenal World as the center. The system interacts with the external environment only indirectly through prediction outcomes, implementing the model dependent ontology idea where the phenomenal world constitutes the agent's reality.

METHODOLOGY

Our methodology centers on systematically embedding the four pillars into an agent and then situating this agent in a controlled grid-based environment that triggers prediction errors at specified intervals. After defining a persistent self-pattern in the form of a small cellular automaton that evolves continuously, the approach incorporates a recurrent module to maintain memory across multiple time steps, ensuring that the agent's internal narrative endures. A parameterized world-model predicts future observations, and whenever the prediction error exceeds a threshold, the agent updates these parameters immediately to adapt in real time rather than relying on offline retraining. The agent also devotes certain steps to generating imaginative or dream-like states via the world-model, effectively decoupling from external inputs and exploring novel representations.

Figure 1 illustrates the architecture, consisting of interconnected modules corresponding to each pillar.

The system can be viewed as an autonomous agent continually interacting with the environment, while an internal loop maintains a coherent sense of self, integrates experiences over time, adapts to unexpected changes, and generates imagined scenarios. Let $x(t)$ denote the sensory inputs at time t, and $a(t)$ the agent action. The agent's internal state is composed of several components: a *self-state* $s(t)$, a *memory state* $h(t)$, and a latent *world-model state* $w(t)$.

Self-Instantiation: Persistent Self-Representation

The self-instantiation pillar endows the agent with an explicit internal representation of itself. We implement this as a recurrent sub-network that maintains a persistent *self-state* vector $s(t)$, which can be thought of as the agent's internal identity at time t. Unlike conventional hidden states that are reset between episodes, $s(t)$ is continuously carried forward and updated. Maintaining a persistent $s(t)$ allows the agent to refer back to "itself" at earlier times, enabling higher-order reasoning and introspection. The decision-making module can query $s(t)$ for consistency checks or to align action selection with long-term goals or identity.

Temporal Continuity: Memory and Narrative Maintenance

Temporal continuity is achieved through mechanisms that ensure the agent's internal state $h(t)$ carries information forward indefinitely, enabling a narrative thread across time. The memory state $h(t)$ is realized by a reservoir of recurrent neurons which accumulates information. The reservoir approach has the advantage of maintaining a fading memory of past inputs while being resistant to catastrophic forgetting since the high-dimensional dynamics are only indirectly adjusted via a trainable readout layer. Importantly, $h(t)$ is never arbitrarily reset during the agent's lifetime, forcing it to confront the consequences of forgetting important information. This leads to behaviors where the agent actively reinforces critical knowledge, creating an internal narrative to keep salient facts accessible.

Disturbance-Driven Adaptation: Learning From Surprise

The third pillar introduces an online learning loop activated by *disturbances* – significant discrepancies between predictions and actual observations. If a predicted error exceeds a threshold, it is flagged as a disturbance. Upon detecting a disturbance, the adaptation module engages a rapid learning process to update relevant parts of the system. We employ a combination of short-term plasticity and meta-learning approaches for these updates. Because $s(t)$ and $h(t)$ preserve context, the adaptation is context-sensitive: the agent effectively "knows" when the surprise occurred, helping attribute causes and adjust appropriate components. This ensures the agent remains robust and responsive over long durations, incorporating new information on the fly rather than waiting for offline retraining.

Autonomous World-Building: Internal Simulation and Imagination

The final pillar enables the agent to construct and explore an internal "world"—a simulated model of its environment and possible futures. This world model is represented by a latent state $w(t)$ which the agent can use to imagine scenarios without external inputs. At each time step, the agent allocates some computation to "dreaming": using the world model to generate hypothetical next states. These simulated sequences provide the agent with counterfactual experiences: it can anticipate outcomes of actions without executing them, or experiment with scenarios that have never occurred. The autonomous aspect implies the agent does this proactively – it has an intrinsic drive to engage in world-building, rather than only using the model when required for immediate decisions. The world-building capability provides a cognitive sandbox for the agent. During periods when external inputs are limited, it can continue to enrich its knowledge by simulating scenarios—analogous to how human creativity often involves mentally simulating hypothetical situations.

Unified Mathematical Formulation

The integration of our four pillars can be formalized as a unified discrete-time update equation system:

$$\begin{cases} \mathrm{h}_{\mathrm{t}\,+\,1} = \mathrm{F}_{\mathrm{self}}\left(\mathrm{h}_{\mathrm{t}},\ \mathrm{a}_{\mathrm{t}}\ \mathrm{x}_{\mathrm{t}\,+\,1},\ \mathrm{w}_{\mathrm{t}}\right), \\ \mathrm{w}_{\mathrm{t}\,+\,1} = \mathrm{F}_{\mathrm{word}}\left(\mathrm{w}_{\mathrm{t}}, \mathrm{a}_{\mathrm{t}}\right) + \mathrm{K}_{1}\left[\mathrm{x}_{\mathrm{t}\,+\,1} - \mathrm{G}_{\mathrm{world}}\left(\mathrm{F}_{\mathrm{world}}\left(\mathrm{w}_{\mathrm{t}}, \mathrm{a}_{\mathrm{t}}\right)\right)\right], \\ \theta_{\mathrm{t}\,+\,1} = \theta_{\mathrm{t}} + \eta\frac{\partial}{\partial\theta}\left[\mathrm{x}_{\mathrm{t+1}} - \mathrm{G}_{\mathrm{world}}\left(\mathrm{F}_{\mathrm{world}}\left(\mathrm{w}_{\mathrm{t}}, \mathrm{a}_{\mathrm{t}}\right)\right)\right]^{2} \end{cases}$$

This system describes how the self-state, world model, and adaptation mechanisms interact. The first equation combines self-instantiation and temporal continuity, updating the agent's internal state based on previous state, action, sensory input, and world model state. The second equation implements world-building with a prediction-correction mechanism similar to a Kalman filter. The third equation represents disturbance-driven adaptation, updating model parameters via gradient descent on prediction error.

Table 1: Summary of the mathematical model. Each equation corresponds to one or more pillars.

Equation	Related Pillars	Function
$\mathbf{h}_{\mathbf{t}\,+\,1}= \mathbf{F}_{\mathrm{self}}\left(\mathbf{h}_{\mathbf{t}},\ \mathbf{a}_{\mathbf{t}}\ \mathbf{x}_{\mathbf{t}\,+\,1}, \mathbf{w}_{\mathbf{t}}\right)$	Self-Instantiation & Temporal Continuity	Maintains a persistent self- representation, ensuring identity across time.
$\mathbf{w}_{\mathbf{t}\,+\,1}=\mathbf{F}_{\mathrm{word}}\left(\mathbf{w}_{\mathbf{t}}, \mathbf{a}_{\mathbf{t}}\right)+ \mathbf{K}_{1}[\mathbf{x}_{\mathbf{t}\,+\,1}-\mathbf{G}_{\mathrm{world}} \left(\mathbf{F}_{\mathrm{world}}\left(\mathbf{w}_{\mathbf{t}},\mathbf{a}_{\mathbf{t}}\right)\right)]$	World-Building & Predictive Learning	Allows the AI to simulate, anticipate, and correct its internal world model based on real-world feedback.

Continued

Table 1: Continued

Equation	Related Pillars	Function
$\theta_{t+1}=\theta_t+\eta\frac{\partial}{\partial\theta}\left[\mathbf{x}_{t+1}-\mathbf{G}_{\text{world}}\left(\mathbf{F}_{\text{world}}\left(\mathbf{w}_t,\mathbf{a}_t\right)\right)\right]^2$	Disturbance-Driven Adaptation	Enables on-the-fly learning when predictions fail, allowing resilient adaptation.

EXPERIMENTAL SETUP

To test our proposed four-pillar framework, we implemented a proof-of-concept simulation that models an autonomous agent operating in a simplified grid-based environment. In this setup, the agent's entire "experience" is generated from its internal model rather than from direct sensory contact with an external world. This design forces the agent to rely solely on its predictive constructs, making it an ideal testbed for evaluating our hypotheses.

Implementation of the Four-Pillar Framework

Our implementation consists of four interconnected modules corresponding to the four pillars:

Self-Instantiation: This module is realized through a cellular automata (CA) system. The CA runs continuously and independently of direct environmental inputs, evolving complex patterns that serve as the agent's persistent self-model. These patterns represent the agent's internal identity and allow it to reference and build upon its past states, ensuring that its self-representation is maintained over time.

Temporal Continuity: To establish a continuous narrative, we implemented a dedicated memory structure that integrates state information across time steps. Unlike conventional neural networks that reset their hidden states, our memory component accumulates sequential data, preserving the context of past experiences. This continuous thread of information enables the agent to form a coherent narrative, linking past, present, and anticipated future states.

Disturbance-Driven Adaptation: Adaptation is driven by prediction errors in our framework. We simulate environmental disturbances by deliberately introducing discrepancies between predicted and actual sensory inputs at predetermined intervals (specifically, at steps 15 and 40). When such disturbances occur, the agent's online learning loop is activated: the system adjusts its parameters proportionally to the magnitude of the prediction error. This real-time learning process allows the agent to adapt to unexpected changes, reflecting its ability to learn from surprise.

Autonomous World-Building: This module is divided into two parts. First, the Internal World Model is continuously updated based on the prediction errors, forming a dynamic simulation of the external environment. Second, the Dream/Prediction Component explores possibilities beyond immediate inputs by generating counterfactual scenarios and "dream sequences." These imaginative extensions enable the agent to go beyond mere replication of reality and develop a richer, internally generated experience.

In our simulation, the agent interacts with a simplified grid-world. Notably, all its actions and learning are based solely on its internal world model; there is no direct, unmediated access to the real environment. This emphasizes the framework's central hypothesis that cognition is entirely an internally constructed phenomenon.

Metrics and Evaluation

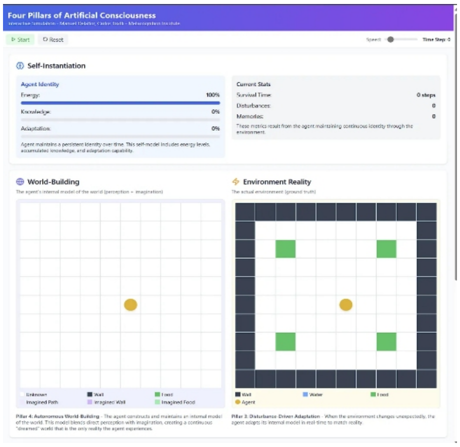
To objectively assess the development and performance of each component, we devised a comprehensive set of metrics. The Self-Instantiation Metric measures the complexity and stability of the CA-generated patterns, serving as an indicator of the robustness of the agent's self-model. A Temporal Continuity Metric quantifies the system's ability to integrate and maintain historical information across time, reflecting the coherence of its internal narrative. The Adaptation Metric evaluates the agent's response to simulated disturbances by measuring the speed and accuracy of its corrective adjustments following prediction errors. A World-Building Metric assesses the originality and coherence of the internal world model and the dream sequences, providing insight into the agent's capacity for imaginative simulation. Finally, an Overall Reference Metric aggregates these metrics to provide a composite measure of the emergent conscious-like behavior of the agent.

These metrics were specifically designed to quantify emergence of consciousness-like properties according to MDO principles. The Self-Instantiation Metric measures pattern persistence and complexity using Shannon entropy and recurrence quantification, capturing the system's ability to maintain a stable self-model despite environmental changes. Temporal Continuity is evaluated through mutual information between time-separated states, quantifying how effectively past experiences inform current processing. The Adaptation Metric combines learning rate and error reduction following disturbances, measuring real-time parameter adjustments. World-Building is assessed through divergence between generated dream sequences and observed environment patterns, using Jensen-Shannon divergence to quantify creative extensions beyond mere replication. Unlike evaluations in purely functionalist frameworks that focus on task performance, these metrics specifically target the system's ability to construct and maintain its own phenomenal models—the core principle distinguishing our MDO-based approach from conventional AI systems.

These metrics are recorded over multiple simulation steps, allowing us to monitor the evolution and interaction of each pillar over time. Correlating the metrics with the introduced disturbances and subsequent adaptations enables us to evaluate the effectiveness of our framework in generating robust, continuous, and adaptive behaviors. This experimental setup provides a rigorous yet transparent method to validate the feasibility of our framework and to quantify the properties of our proposed artificial consciousness.

RESULTS

The evolution of the preliminary proof of concept (PoC) was evaluated over a 75-step simulation, and the outcomes demonstrate the gradual emergence of cognitive-like behavior. In the initial stage the internal world model is empty, indicating that no autonomous construction is present at the outset. By step 15 we can see that the internal world model is starting to mimic the environment, suggesting that the processes underlying autonomous world-building and internal adaptation are taking hold. By step 45 the autonomous construction is finished, representing the "world" the agent is interacting with, and we can see its dreaming sequence displaying novel elements not found in the actual environment. This divergence provides compelling evidence that the system is not merely replicating its sensory inputs but is constructing an internal, imaginative narrative.



Time Step: 0

Agent Identity

Energy: 100%

Knowledge: 0%

Adaptation: 0%

Agent maintains a persistent identity over time. This self-model includes energy levels, accumulated knowledge, and adaptation capability.

Current Stats

Survival Time: 0 steps

Disturbances: 0

Memories: 0

These metrics result from the agent maintaining continuous identity through the environment.

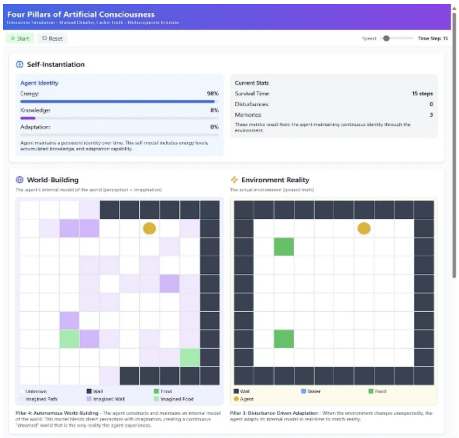
The agent's internal model of the world (perception + imagination)

The actual environment (ground truth)

Pillar 4: Autonomous World-Building - The agent constructs and maintains an internal model of the world. This model blends direct perception with imagination, creating a continuous "dreamed" world that is the only reality the agent experiences.

Pillar 3: Disturbance-Driven Adaptation - When the environment changes unexpectedly, the agent adapts its internal model in real-time to match reality.

Figure 2: Initial state (step 0) showing environment and internal world model. At this early stage, the internal model is empty, demonstrating minimal autonomous construction.



Time Step: 15

Agent Identity

Energy: 98%

Knowledge: 8%

Adaptation: 0%

Agent maintains a persistent identity over time. This self-model includes energy levels, accumulated knowledge, and adaptation capability.

Current Stats

Survival Time: 15 steps

Disturbances: 0

Memories: 3

These metrics result from the agent maintaining continuous identity through the environment.

The agent's internal model of the world (perception + imagination)

The actual environment (ground truth)

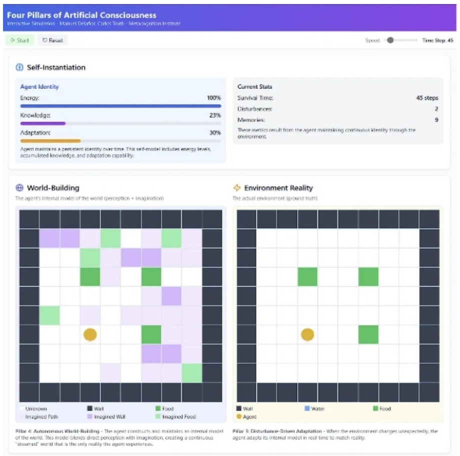
Pillar 4: Autonomous World-Building - The agent constructs and maintains an internal model of the world. This model blends direct perception with imagination, creating a continuous "dreamed" world that is the only reality the agent experiences.

Pillar 3: Disturbance-Driven Adaptation - When the environment changes unexpectedly, the agent adapts its internal model in real-time to match reality.

Figure 3: Step 15 in simulation. By this point, the internal model has begun to develop to adapt its behavior to the environment.

Four-Pillar Visualization

Figure 4 provides a visualization of all four framework components at step 45. This figure illustrates the interplay between the cellular automata (representing Self-Instantiation), the internal world model (demonstrating World-Building), the dream sequence (illustrating creative extensions), and the world model confidence (indicating successful adaptation). The integrated visualization underscores how the four pillars collectively contribute to emergent behavior.


Four Pillars of Artificial Consciousness
Interactive Simulation - Manuel Delaflor, Carlos Toxtli - Metacognition Institute

Time Step: 45

Self-Instantiation

Agent Identity

Energy:	100%
Knowledge:	23%
Adaptation:	30%

Agent maintains a persistent identity over time. This self-model includes energy levels, accumulated knowledge, and adaptation capability.

Current Stats

Survival Time:	45 steps
Disturbances:	2
Memories:	9

These metrics result from the agent maintaining continuous identity through the environment.

World-Building

The agent's internal model of the world (perception + imagination)

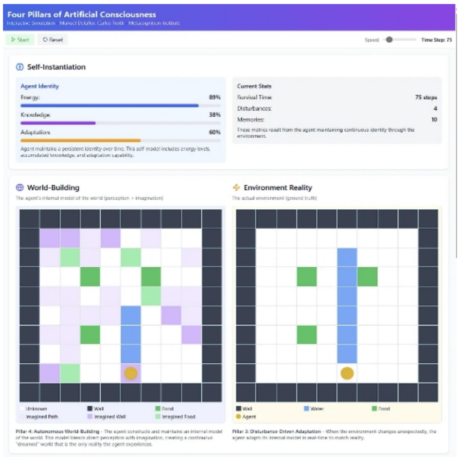
Pillar 4: Autonomous World-Building - The agent constructs and maintains an internal model of the world. This model blends direct perception with imagination, creating a continuous "dreamed" world that is the only reality the agent experiences.

Environment Reality

The actual environment (ground truth)

Pillar 3: Disturbance-Driven Adaptation - When the environment changes unexpectedly, the agent adapts its internal model in real-time to match reality.

Figure 4: Visualization of all four framework components at step 45: (A) Cellular automata patterns representing Self-Instantiation; (B) Internal world model demonstrating World-Building; (C) Dream sequence illustrating autonomous imagination; and (D) Adaptation confidence map showing system response to disturbances.


Four Pillars of Artificial Consciousness
Interactive Simulation - Manuel Delaflor, Carlos Toxtli - Metacognition Institute

Time Step: 75

Self-Instantiation

Agent Identity

Energy:	89%
Knowledge:	38%
Adaptation:	60%

Agent maintains a persistent identity over time. This self-model includes energy levels, accumulated knowledge, and adaptation capability.

Current Stats

Survival Time:	75 steps
Disturbances:	4
Memories:	10

These metrics result from the agent maintaining continuous identity through the environment.

World-Building

The agent's internal model of the world (perception + imagination)

Pillar 4: Autonomous World-Building - The agent constructs and maintains an internal model of the world. This model blends direct perception with imagination, creating a continuous "dreamed" world that is the only reality the agent experiences.

Environment Reality

The actual environment (ground truth)

Pillar 3: Disturbance-Driven Adaptation - When the environment changes unexpectedly, the agent adapts its internal model in real-time to match reality.

Figure 5: Final state (step 75) an unexpected disturbance appears on the environment, and the agent adapts, demonstrating autonomous world-building capabilities.

Component Development Metrics

Figure 6 displays the progression of metrics over the entire 75-step simulation, providing a quantitative overview of how each pillar of the framework evolves. In this graph, the x-axis represents the simulation steps (from 0 to 75), while the y-axis shows the normalized metric values (0-1 scale). The plot includes color-coded curves for each pillar: self-instantiation (blue), temporal continuity (green), adaptation (red, with notable responses at disturbance points marked at steps 15 and 40), and world-building (purple), along with an overall reference metric (black).

Component Results

The experimental data clearly indicate that the integrated framework yields robust, emergent cognitive behavior. At the outset, the self-instantiation component exhibits a modest score of approximately 0.11, which steadily increases to around 0.60 by the end of the 75-step simulation, reflecting the system's growing ability to autonomously maintain its internal processes. Concurrently, the temporal continuity metric shows a pronounced improvement, peaking near 0.60 around step 50, thereby evidencing the system's effective integration of sequential experiences into a coherent internal narrative. In response to controlled disturbances, the adaptation component reveals distinct step-wise increases—beginning at a baseline of about 0.10 and climbing to roughly 0.20—with marked jumps at steps 15 and 40 that coincide precisely with the induction of environmental perturbations. Notably, the world-building component develops from an initial value of around 0.11 to approximately 0.34 over the course of the simulation, signifying that the agent is not merely replicating sensory inputs but is actively constructing an internal "world" characterized by a unique dreaming sequence that introduces novel elements absent in the actual environment. The cumulative effect of these individual processes is reflected in the overall reference metric, which increases steadily from approximately 0.11 to 0.44, thereby reinforcing the thesis that sentience-like properties can emerge from the integration of self-instantiation, temporal continuity, disturbance-driven adaptation, and autonomous world-building.

Component Interactions and Limitations

Self-instantiation appears to provide a foundation upon which temporal continuity can build, as evidenced by similar growth patterns. The adaptation component shows discrete jumps in response to disturbances, while world-building develops more gradually as the system accumulates experience. Current limitations of this first approach include: (1) the initial internal world model closely resembles the environment, suggesting some initial copying rather than purely autonomous construction; (2) a simplified grid-based environment rather than more complex scenarios; and (3) relatively small-scale neural networks and cellular automata, limiting the complexity of emergent behaviors. These results are consistent with our hypothesis that current AI approaches focusing solely on pattern recognition miss essential ingredients for consciousness-like properties, and that the world-building aspect creates a genuinely autonomous phenomenal experience. As a PoC, these preliminary findings suggest directions for more extensive investigations.

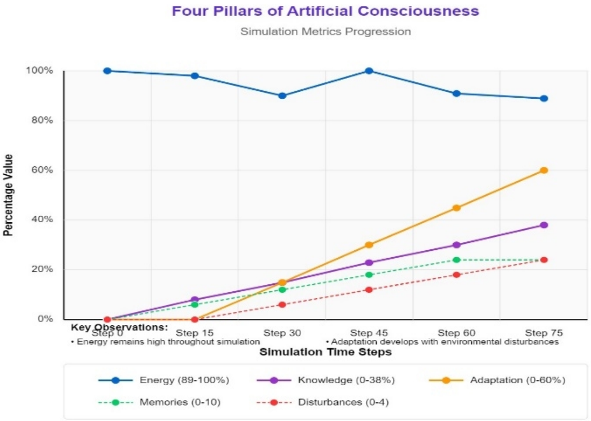


Figure 6: Development of metrics over 75 simulation steps. The graph shows the progression of all four pillars: self-instantiation (increasing from 0.11 to 0.60), temporal continuity (peaking at 0.60 around step 50), disturbance-driven adaptation (showing distinct jumps at steps 15 and 40, marked with vertical dotted lines), and world-building (growing from 0.11 to 0.34). The overall reference metric (black line) demonstrates the emergent properties from the integration of all components.

DISCUSSION

This section revisits the research questions; for RQ1, self-instantiation fosters stable morphological patterns in the cellular automaton. RQ2 is advanced by continuous memory, which preserves context over time to enhance coherence. Disturbance-driven adaptation answers RQ3, outperforming offline learning but complicating parameter management due to rapid updates. Finally, dream phases for autonomous world-building illuminate RQ4, as they enrich representations but raise alignment concerns if emergent goals arise. The results indicate strong potential for robust, real-time learning alongside careful oversight of creative internal processes. The experimental results are consistent with our hypothesis that integrating self-modeling, memory, adaptation, and imagination yields more robust and flexible cognition. While we use terms like “self-awareness” or “imagination,” we acknowledge that the agent’s subjective experience (if any) remains unknown—we demonstrate functional analogues of conscious processes. One strength of our approach is that it marries theoretical concepts with practical implementation. Even partially endowing AI with conscious-like features yields tangible benefits for performance and adaptability. This has implications for AI safety: a system that monitors itself and adapts may avoid certain failure modes autonomously.

Our framework builds upon and extends a rich body of literature in machine consciousness and cognitive architectures. Foundational models such as Global Workspace Theory (Baars, 1988b) and Integrated Information Theory (Oizumi et al., 2014b; Tononi, 2004b) provide important insights into the structure of consciousness, while more recent efforts like Ha and Schmidhuber's world models (Oizumi et al., 2014b; Tononi, 2004b) and Bengio's Consciousness Prior (Bengio, 2017a) have advanced our understanding of internal representation learning. In addition, research on curiosity-driven exploration (Pathak et al., 2017a; Schmidhuber, 1991) and studies of mental representations in machine learning (Butlin, 2023) underscore the importance of adaptive mechanisms for robust cognitive performance. They complement our approach by highlighting the necessity of continual adaptation and internal simulation for emergent behavior.

Challenges remain, including the reliability of the world-building component—if imagined scenarios diverge too much from reality, it could lead to suboptimal decisions—and the scalability to higher-dimensional inputs or more complex cognitive tasks. Moreover, the emergence of agents with conscious-like properties raises ethical considerations. If an AI system displays behaviors associated with consciousness, even in a rudimentary form, it prompts questions regarding their ethical treatment. Our framework could also serve as a research tool in cognitive science, testing theories about the minimal conditions for consciousness by examining which combinations of pillars yield which behaviors. Ethical implications of such systems have also been discussed in works by (Chalmers, 1996) and (Wallach and Allen, 2008), emphasizing the broader societal impact of developing AI with advanced cognitive features. Future research directions include enhancing each pillar (e.g., giving the agent a more complex self-model), testing the framework in games, for instance, fabricating and NPC to interact in a previously designed environment like Minecraft, embodied robotics or social agents, exploring metrics for consciousness in machines, and investigating the ethical implications of increasingly autonomous and life-like AI.

CONCLUSION

We presented a novel four-pillar framework for developing AI systems with emergent conscious-like behavior, integrating self-instantiation, temporal continuity, disturbance-driven adaptation, and autonomous world-building. This approach addresses key limitations of conventional AI: the lack of a persistent self, inability to maintain long-term context, rigidity in the face of change, and absence of imagination. Our proof-of-concept implementation demonstrates the feasibility of operationalizing these principles and showed promising development across all components. It is important to emphasize that this work represents an early exploration of these ideas rather than a complete solution to machine consciousness. Many challenges remain, including scaling the architecture into more complex environments, developing more sophisticated metrics for evaluating conscious-like properties, and addressing ethical implications of creating increasingly autonomous agents. We believe that the four-pillar framework demonstrates that key ingredients associated with consciousness can be built into AI systems today, yielding benefits in adaptability and behavioral coherence. This opens a pathway for interdisciplinary exploration of consciousness in artificial entities, moving from theoretical postulates to implemented systems and bringing us closer to AI that possesses a rich internal life guiding its intelligence in human-like ways.

REFERENCES

Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, 7055–7066.

Bernard J. Baars. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, UK.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences* 40 (2017), 1–72. doi: 10.1017/S0140525X16001837.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proceedings of the 34th International Conference on Machine Learning* (2017), 1126–1135.

David Gamez. 2008. Progress in Machine Consciousness. *Consciousness and Cognition* 17, 3 (2008), 887–910.

David Ha and Jürgen Schmidhuber. 2018. World Models. In *Advances in Neural Information Processing Systems*.

David J. Chalmers. 2023. Could a Large Language Model Be Conscious? *Journal of Consciousness Studies* 30, 3–4 (2023), 10–45. doi: 10.56203/JCS.2023.2.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-Driven Exploration by Self-Supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Sydney, Australia, 2778–2787.

Giulio Tononi. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5, 1 (2004), 42. doi: 10.1186/1471-2202-5-42.

Igor Aleksander. 2007. *How to Build a Mind: Toward Machines with Imagination*. Columbia University Press.

Jürgen Schmidhuber. 1991. A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers. In *Proceedings of the International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, MA, 222–227.

James A. Reggia. 2013. The Rise of Machine Consciousness: Studying Consciousness with Computational Models. *Neural Networks* 44 (2013), 112–131. doi: 10.1016/j.neunet.2013.03.011.

Mantas Lukoševičius and Herbert Jaeger. 2009. Reservoir Computing Approaches to Recurrent Neural Network Training. *Computer Science Review* 3, 3 (2009), 127–149. doi: 10.1016/j.cosrev.2009.03.005.

Manuel Delaflor. 2024. *Introduction to MDO*. Technical Report. Metacognition Institute. doi: 10.13140/RG.2.2.16809.20325 Licensed under CC BY-NC-ND 4.0.

Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology* 10, 5 (2014), e1003588. doi: 10.1371/journal.pcbi.1003588.

Patrick Butlin. 2023. Mental Representations and Machine Learning. *Minds and Machines* 33, 2 (2023), 197–218. doi: 10.1007/s11023-022-09602-0 David J. Chalmers. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, NY. 1–414 pages.

Stanislas Dehaene. 2017. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. In *Proceedings of the Royal Society of London*.

Wendell Wallach and Colin Allen. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Yoshua Bengio. 2017. The Consciousness Prior. arXiv preprint arXiv:1709.08568. arXiv:1709.08568 [cs. AI] pp. 1–8.