ISSN 2319 – 7595
Volume 15, No.1, January – February 2026
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs011512026.pdf
https://doi.org/10.30534/ijiscs/2026/011512026

Synthetic Consciousness Foundation for Responsible and Accountable AI Agents

[1]Muni Madhuri Koilakunta, [1]Henry Hexmoor
[1]College of Computing, Southern Illinois University, Carbondale, IL 62901
[1]munimadhuri.koilakuntla@siu.edu, hexmoor@siu.edu

Received Date : November 18, 2025 Accepted Date : December 22, 2025 Published Date : January 07, 2026

ABSTRACT

Agentic AI aspires to act autonomously, reason purposefully, and interact meaningfully with others (agents or persons). These agents must proactively adhere to constitutional rubrics in their environment hence be responsible. Constitutional rubrics for agents are structured sets of principles and evaluation criteria that define, operationalize, and continuously check the "constitution" (values, constraints, and behaviors) of autonomous AI agents. They sit between high-level AI constitutions (e.g., safety or ethical principles) and low-level policies (e.g., allow/deny rules), giving you something concrete you can implement, score, and audit.

Key words: Agents AI, Synthetic Consciousness, Cognitive Modeling, Artificial Intelligence

1. INTRODUCTION

Agentic AI aspires to act autonomously, reason purposefully, and interact meaningfully with others (agents or persons). These agents must proactively adhere to *constitutional rubrics* in their environment hence be *responsible*. Constitutional rubrics for agents are structured sets of principles and evaluation criteria that define, operationalize, and continuously check the "constitution" (values, constraints, and behaviors) of autonomous AI agents. They sit between high-level AI constitutions (e.g., safety or ethical principles) and low-level policies (e.g., allow/deny rules), giving you something concrete you can implement, score, and audit.

Consequences of their actions must identify them as originators of actions, hence be *accountable*. Responsibility and accountability predicate a form of synthetic consciousness (SC) [1][2][3][13][14][15]. We endeavor to formulate SC to lay the foundation for agentic responsibility and accountability. We post it that there is no requirement to understand or to compare SC with biological consciousness that largely rely on sentience and sapience.

Artificial intelligence (AI) has evolved from rule-based systems to advanced deep learning networks that can recognize images, translate languages, and even generate human-like text. Yet, despite these impressive capabilities, most AI systems remain narrow in scope. They perform tasks efficiently but lack a deeper understanding of what they are doing. These systems follow learned patterns without awareness, intention, or the ability to reflect on their own behavior. This difference between high performance and true understanding continues to separate artificial intelligence from human-like cognition. Humans, on the other hand, learn through experience, attention, and awareness. We don't just react to the world, we interpret it, predict outcomes, and adjust our actions based on goals and internal states. For instance, when we feel hungry, anxious, or curious, these emotions influence how we behave and what we focus on. Such processes involve not just learning patterns but also maintaining an internal model of the world and ourselves within it. Replicating this kind of adaptive, self-regulating intelligence is one of the biggest challenges in modern AI research. Currently, we are now exploring the intersection of neuroscience and artificial intelligence, seeking inspiration from how the brain organizes thought, emotion, and awareness. The human brain is not a single processor but a network of specialized regions constantly exchanging information. Yet, at any given moment, only a small portion of this information becomes conscious: the part that reaches our awareness and guides our decision-making. Understanding how this selection happens offers valuable clues for designing more intelligent, autonomous agents. The field of cognitive architectures attempts to build such systems agents that can perceive, reason, plan, and act in a coordinated manner. Some well-known architectures like SOAR, ACT-R, and LIDA have modelled aspects of human cognition such as memory, reasoning, and goal management. However, many of these still depend on predefined rules or symbolic reasoning, which limits their flexibility in dynamic, uncertain environments. There remains a need for architectures that can not only process information but also learn how to learn, updating internal representations continuously and adjusting behavior in response to changing goals and conditions. This project explores one such direction through the concept of synthetic consciousness, the attempt to simulate conscious-like processes in artificial systems. The goal is not to claim true human consciousness, but to reproduce some of its core functional properties: awareness, attention, adaptability, and emotional regulation. The proposed system is implemented in the MiniGrid environment, where the agent must navigate and retrieve objects under limited visibility and changing conditions. It learns by predicting what will happen next,

comparing its predictions to sensory feedback, and adjusting its internal model to minimize errors, much like how the brain continuously refines its expectations of the world. We are striving to integrate two major theoretical frameworks from cognitive neuroscience: Global Workspace Theory (GWT) and Predictive Processing (PP) [8][9][10][11]. GWT explains how information becomes globally available across different cognitive modules, enabling attention and decision-making, while PP provides a mechanism for learning and adaptation through prediction-error minimization. Together, they form the foundation for building a self-organizing, goal-driven agent that learns to act intelligently in a dynamic world.

Emotions and internal drives such as energy, threat, and curiosity are modelled as control signals that influence the agent's decisions. This objective aims to simulate emotional modulation of behavior, ensuring that the agent maintains stability between exploration and safety, ultimately achieving adaptive balance in changing environments.

Intelligent agents must operate under uncertainty, partial observability, and constantly changing goals. They must prioritize attention, balance multiple internal needs, and adapt to new situations without external supervision. However, most existing reinforcement learning and deep-learning architectures are limited by their narrow focus on input–output optimization. They struggle to generalize across different contexts, fail to retain useful memory of past experiences, and cannot dynamically coordinate competing objectives such as safety, curiosity, or energy conservation. This results in brittle and inefficient behavior when confronted with unfamiliar patterns or conflicting goals [4][5][6]. A second major challenge lies in the absence of global coordination and interpretability within current neural systems. Traditional deep networks distribute processing across layers but lack a central mechanism to integrate and broadcast relevant information to all subsystems, a hallmark of conscious cognition in biological systems. As a result, such models often suffer from fragmented perception, slow adaptation, and limited transparency in decision-making. Moreover, emotional or motivational factors, which play a vital role in biological learning and attention, are typically absent from artificial agents, leading to less natural, rigid patterns of behavior. This work aims to bridge the gap between purely algorithmic intelligence and biologically inspired cognition by combining Global Workspace Theory (GWT) and Predictive Processing (PP) into a single framework for synthetic consciousness [2]. Next, we describe design of an agent. At each step, the agent observes the environment, predicts the next sensory input, evaluates the difference (prediction error), adjusts internal emotional states, and selects the optimal action using the free-energy minimization principle [11].

2. AGENT ARCHITECTURE

1. *Initia*lize environment E (MiniGrid)
2. Initialize model parameters θ for all modules:
Sensory, Memory, Emotion, Workspace, and Policy
3. Initialize replay buffer B = ∅
4. Set initial emotional states (energy, threat, curiosity)
5. for each episode = 1 to N do
6. Reset environment and obtain initial observation s0
7. Initialize episode reward R = 0
8. for each timestep t = 1 to T do
--- Perception and Prediction ---
9. Encode sensory input xt = SensoryModule(st)
10. Predict next state ŝt+1 = PredictiveModel(xt)
11. Compute prediction error εt = |st+1 - ŝt+1|

--- Global Broadcasting ---
12. if εt > threshold or high emotional salience then
13. Broadcast salient information via GlobalWorkspace
14. Update working memory and attention focus
15. end if
--- Emotion and Homeostasis Update ---
Update internal drives:
16. energy ← energy - cost(action)
17. threat ← evaluate_environment(st)
18. curiosity ← novelty_score(st)
19. Compute homeostatic deviation H
--- Action Selection via Free-Energy Minimization ---
20. For each possible action a in A:
21. FE(a) = εt + wE*(energy_target - energy)+ wT*(threat - threat_target)+ wG*(1 - goal_progress)
22. Choose action at = argmin(FE(a)) with ε-greedy exploration
--- Environment Interaction ---
23. Execute at, receive reward rt and next state st+1
24. Store transition (st, at, rt, st+1, εt) in buffer B
25. Update R ← R + rt
--- Learning Step ---
26. Sample mini-batch from B
27. Update model parameters θ to minimize:
28. L = α*(prediction error) + β*(free-energy cost)
29. end for
Log episode metrics: total reward, average ε, entropy, H
30. end for

3. IMPLEMENTATION

The proposed system was implemented through a modular architecture in which perception, prediction, attention, emotion, and action selection operate as coordinated but functionally distinct subsystems. Each module is instantiated as a trainable component, and together they form a closed-loop cognitive cycle that governs the agent's adaptive behavior. The MiniGrid environment was selected due to its constrained observability, sparse rewards, and dynamic layouts, which provide an ideal testing ground for evaluating synthetic consciousness driven attention and prediction mechanisms. Figure 1 shows a snapshot of the Minigrid agent environment. The major components are outlined next.

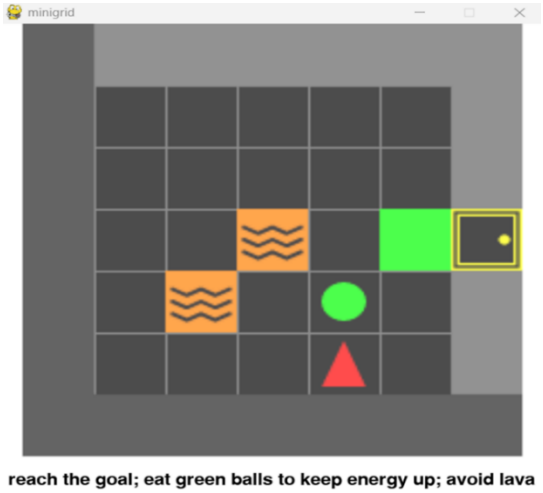

Figure 1: The MiniGrid Environment

A. Sensory Processing and State Encoding

The sensory module receives raw grid observations from MiniGrid, consisting of spatially structured RGB encodings and agent-relative orientation cues. These inputs are passed through a lightweight convolutional encoder implemented in PyTorch, compressing the environment state into a latent feature representation. This latent state serves two purposes:

1. It provides a compact summary of the environment necessary for efficient predictive modeling, and
2. It forms the perceptual content eligible for global broadcasting within the workspace.

   By encoding only the most relevant perceptual attributes, the sensory system minimizes input noise and ensures that downstream modules operate on stable, interpretable state vectors.

B. Predictive Processing and Forward Modeling

At each timestep, the predictive model generates a forecast of the agent's next sensory state given its current perception and hypothesized action. Implemented as a recurrent neural architecture, the model learns a state-transition distribution mapping latent states to expected future states. Prediction error is computed as the absolute difference between the predicted sensory embedding and the actual sensory feedback provided by the environment. This prediction error signal serves as the primary driver of agent learning, modulating attention, memory consolidation, and emotional homeostasis. High prediction error marks unexpected changes or failures of expectation, causing the agent to re-evaluate internal beliefs, broadcast salient discrepancies to the global workspace, and adjust its policy accordingly[7].

C. Global Workspace Broadcasting and Attention Control

The global workspace serves as the coordination center for all cognitive activity. Incoming perceptual or emotional signals must compete for access to the workspace through a salience-driven gating mechanism inspired by Global Workspace Theory. When prediction error surpasses a threshold, or when emotional variables signal urgency (e.g., elevated threat), the corresponding informational fragment is globally broadcast. This broadcast synchronizes all modules, policy selection, memory retrieval, and emotion regulation, by making the salient content immediately accessible across the entire architecture. In practice, this is implemented as a shared tensor buffer whose contents are overwritten only when new, high-salience information emerges. The workspace thus functions as both an attentional bottleneck and an integrative meta-controller that ensures coherent, unified decision-making rather than fragmented subsystem behavior.

D. Emotion and Homeostasis Regulation

Emotional modeling was crucial for generating adaptive, human-like modulation of behavior. The system tracks three core drives: energy, threat, and curiosity. Energy acts as a fatigue-like depletion variable that penalizes unnecessary movement. Threat evaluates environmental risk using distance to obstacles, unknown regions, or adversarial cues. Curiosity measures novelty and encourages exploration of previously unseen states. These drives evolve dynamically based on the agent's sensory feedback and internal predictions. They directly influence the free-energy cost function, allowing emotional states to bias decision-making toward safety, efficient resource use, or exploratory behavior when appropriate. This coupling between emotion and action selection produces more stable and context-responsive behavior than reward-driven reinforcement learning alone.

E. Policy Module and Free-Energy Minimization

Action selection is governed by a free-energy optimization process in which candidate actions are ranked according to predicted prediction error, emotional deviation, and goal progression. Instead of selecting actions solely for extrinsic reward, the agent evaluates how each action will shape its internal predictive certainty and its alignment with homeostatic goals. This framework enables the agent to self-organize its behavior toward reducing uncertainty while maintaining emotional stability, representing a functional analogue of conscious deliberation.

F. Memory, Replay Buffer, and Learning Dynamics

All sensory states, actions, rewards, and prediction errors are stored in a replay buffer for experience-driven learning. During training, mini-batches of transitions are sampled to update both the predictive model and the policy module. The loss function integrates prediction error, free-energy deviation, and homeostatic stability cost, allowing the system to simultaneously refine perceptual expectations and behavioral strategies. This memory-driven learning mechanism ensures that the agent not only reacts to immediate stimuli but also gradually develops structured internal models of its environment.

G. Emergent Behavior and System Dynamics

Across training episodes, the integration of predictive processing with global broadcasting yields observable emergent patterns. The agent begins to allocate attention efficiently, suppressing irrelevant sensory signals and focusing on elements critical for task completion. Emotional regulation reduces erratic movement and encourages balanced exploration. Workspace entropy stabilizes as the agent learns which sensory signals require global processing and which can remain peripheral. These emergent behaviors such as stability, focused attention, uncertainty reduction, and context-sensitive action, constitute the core functional signatures of synthetic consciousness within this simulation.

Synthetic consciousness directs agent attention. Typical resulting trajectory paths are shown in Figure 2. Guided attention generated by synthetic consciousness substantially lowers error rates of path selection and improves path finding efficiency. The agent outperforms reinforcement learning strategy. Emotional feedback smooths learning and reduces erratic actions. Together, these behaviors represent emergent awareness and adaptive control, i.e., the agent exhibits hallmarks of synthetic consciousness. The system displays emergent traits analogous to awareness and self-control. Predictive Processing reduces uncertainty, while GWT ensures information focus. The agent learns to anticipate, select, and regulate, key markers of synthetic consciousness.

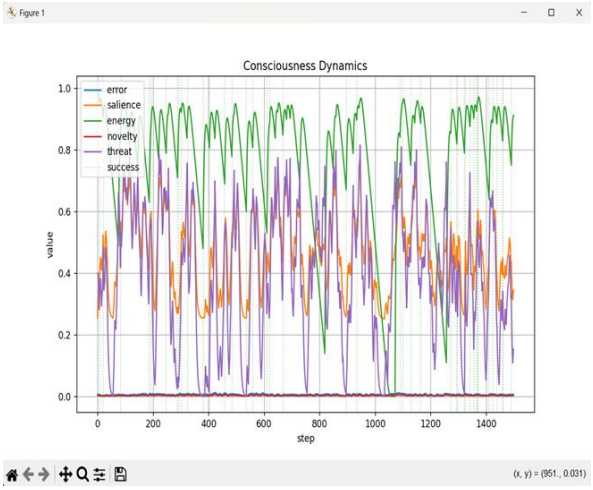

Figure2. Typical Trajectory paths

4. CONCLUSION

In order to endow agents with responsibility and accountability we strive to incorporate a cognitively inspired model of synthetic consciousness. We proposed and implemented the global workspace–Predictive Processing (GWT-PP) architecture that is a hybrid framework that unifies attention, memory, emotion, and predictive learning to simulate aspects of synthetic consciousness.

By combining *global workspace theory*, which governs global attention and information broadcasting, with Predictive Processing, which minimizes sensory uncertainty, the agent demonstrates emergent patterns of self-regulation, adaptability, and goal-directed awareness.

The system was evaluated in the MiniGrid environment using a structured learning pipeline. Quantitative results showed substantial reductions in prediction error, increased reward accumulation, and stable workspace entropy, indicating coherent attentional dynamics. Qualitative analyses revealed emotionally balanced and context-aware decision-making, where the agent dynamically adjusted its actions based on both external goals and internal states. Synthetic consciousness is not about replicating the human mind but about understanding how awareness and adaptation can emerge from computation.

REFERENCES

[1]. Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.

[2]. Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The Global Neuronal Workspace Model: Conscious Access, Attention, and Working Memory. Trends in Cognitive Sciences, 15(6), 292-300. ResearchGate+3PubMed+3SpringerLink+3.

[3]. Baars, B. J. (1997). In the Theatre of Consciousness: Global Workspace Theory, A Rigorous Scientific Theory of Consciousness. Journal of Consciousness Studies.

[4]. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138. Nature+1

[5]. Hohwy, J. (2013). The Predictive Mind. Oxford University Press.

[6]. Clark, A. (2016). Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press.

[7]. Piekarski, M. (2021). Understanding predictive processing: A review. Philosophy Study, 11(9), 555-575. ResearchGate

[8]. Wiese, W., & Metzinger, T. (2017). Bayesian theories of consciousness: A review in search for a minimal unifying model. Neuroscience of Consciousness, 3(1), niw019. PMC

[9]. Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron, 105(5), 776-798.

[10]. Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1668), 20140167.

[11]. Küçükoğlu, B. (2024). Efficient Deep Reinforcement Learning with Predictive Processing (P4O). (Preprint/Article) NBDT Scholastica.

[12]. Chandra Kavi, P., Zamora López, G., & Friedman, D. A. (2024). From Neuronal Packets to Thoughtseeds: A Hierarchical Model of Embodied Cognition in the Global Workspace. arXiv.

[13]. Parr, T., Da Costa, L., & Friston, K. (2020). Neural dynamics under active inference: Plausibility and efficiency of information processing. arXiv preprint. arxiv.org

[14]. B. J. Baars & K. A. McGovern (eds.). (2022). Exploring Evidence for Widespread Integration and Broadcasting in the Brain: Global Workspace Theory.

[15]. Farisco, M. (2024). Is Artificial Consciousness Achievable? Lessons from the Evolution of the Human Brain and Its Relation to AI. Neuroscience & Biobehavioral Reviews. ScienceDirect.