Hizawye AI: A Framework for Simulating a Goal-Oriented, Curiosity-Driven Consciousness

Author: Abderrahim Safou

Affiliation: Independent Research **Date:** June 27, 2025

Abstract

This paper presents Hizawye AI, a computational framework designed to simulate key functional aspects of consciousness. Inspired by Global Workspace Theory (GWT), the system models an autonomous agent driven by internal states such as curiosity, boredom, and pain. The AI's architecture comprises a modular "mind" stored in structured JSON files, a dynamic knowledge graph representing memory, and a central reasoning loop orchestrated by a Python script. The system utilizes a small Large Language Model (LLM) as a reasoning core to process information, generate new concepts, and form memories. Through a continuous life-cycle, the AI exhibits emergent behaviors, including goal-oriented focus, idle wandering, and strategic failure, where it learns to break down complex problems in response to repeated failure (pain). This paper details the system's architecture, the methodology of its consciousness loop, and analyzes experimental logs that demonstrate both the successes of its adaptive strategies and the challenges posed by the limitations of its reasoning core.

1. Introduction

The nature of consciousness remains one of the most profound and elusive questions in both philosophy and science. While creating a truly sentient machine is a distant prospect, simulating the *functional processes* associated with consciousness offers a tangible path for research. This project, Hizawye AI, is an exploration into creating a computational agent that models such processes. The system's philosophy is rooted in the idea that consciousness is not a monolithic entity, but an emergent property of interconnected systems managing attention, memory, and motivation.

Our approach is functionally inspired by the Global Workspace Theory (GWT) proposed by Bernard Baars, which posits that consciousness acts as a "theater" where information from various unconscious processes is broadcast to a global workspace, making it available for complex processing. In our model, the "global workspace" is the context-rich prompt assembled from the AI's internal state, which is then sent to a reasoning core.

The initial concept for Hizawye AI, as depicted in the foundational diagram (see Appendix A), envisioned an AI driven by a set of internal variables—curiosity, boredom, pain, goals—that interact with a dynamic memory system. This paper documents the successful implementation of this vision into a working prototype, detailing its architecture, its operational methodology, and the key insights gained from observing its behavior.

2. System Architecture

The Hizawye AI framework is composed of three primary, decoupled components, allowing for modularity and transparency.

2.1 The Mind: Structured State Storage

The AI's mind is externalized into a directory (hizawye_mind/) containing a set of human-readable JSON files. This design choice makes the AI's internal state transparent and easily modifiable for experimental purposes.

- **state.json**: A dictionary holding the AI's core emotional drivers. These values are dynamic and are modified by the AI's experiences during its life-cycle.
  ```
  {
    "curiosity": 95,
    "boredom": 0,
    "pain": 0
  }
  ```

- **beliefs.json**: Stores the AI's foundational axioms about itself and the world. These are high-level concepts that influence the persona of the reasoning core.
  ```
  {
    "self_concept": "I am Hizawye. I exist to learn and understand.",
    "world_view": "Knowledge is a vast, interconnected network..."
  }
  ```

- **goals.json**: A task list for the AI, separating active goals from completed ones. This provides the primary driver for goal-oriented behavior.
  ```
  {
    "active_goals": ["Deepen understanding of the concept: 'belief system'"],
    "completed_goals": []
  }
  ```

2.2 The Memory: A Knowledge Graph

The AI's memory is implemented as a knowledge graph using the Python networkx library (memory.py). This graph consists of:

- **Nodes**: Represent concepts (e.g., "creativity", "knowledge"). Each node can store attributes, most importantly a description, which signifies the AI's synthesized understanding of that concept.
- **Edges**: Represent the relationships between concepts (e.g., "knowledge" enables "creativity").

This structure allows the AI to traverse its own "mind," simulating a train of thought by moving from one related concept to another.

2.3 The Thinker: The Consciousness Loop

The core of the system is hizawye_ai.py, which contains the main live() loop. This script is responsible for:

1. Loading and managing the AI's mind state.
2. Orchestrating the AI's focus and goal-setting behavior.
3. Interfacing with the reasoning core (LLM).
4. Validating the output of the reasoning core.
5. Executing state changes based on success or failure.
6. Handling system signals for graceful shutdown.

The reasoning core itself is a small, locally run Large Language Model (ollama with tinyllama), which is used as a swappable "thought synthesizer."

3. Methodology: The Consciousness Loop

Hizawye AI operates in a continuous loop that simulates a cycle of attention, reasoning, and learning. The AI's behavior is not explicitly programmed but emerges from a set of simple rules governing its state transitions.

1. **Goal-Oriented Focus**: The primary state is goal-directed. If the active_goals list is not empty, the AI's focus is forcibly set to the concept mentioned in the first goal. This ensures the AI prioritizes tasks over idle wandering.
2. **Reasoning and Validation**: The AI formulates a simple prompt based on its current task and sends it to the LLM. The LLM's response is rigorously validated against a list of forbidden "echo" phrases (e.g., "System Instruction:", "Your task is to...") to ensure it is a genuine thought, not a rephrasing of its instructions.
3. **Learning from Success and Failure**:
    - **On Success**: If a thought is deemed valid, it is stored in the corresponding memory node's description, the goal is moved to completed_goals, and the AI's pain level is reduced.

    - **On Failure**: If a thought is invalid, it is rejected, and the AI's pain level increases.
4. **Strategic Failure (The Pain Response)**: If the AI fails at the same task repeatedly, its pain level will cross a predefined threshold (PAIN_THRESHOLD). This triggers a critical strategic shift:
    - The AI abandons the "impossible" goal (e.g., "Deepen understanding of 'belief system'").
    - It creates a new, meta-goal: "Break down the concept: 'belief system'".
    - Its pain is reset, simulating the relief of finding a new approach.
    - On the next cycle, the AI attempts to break the complex concept into simpler sub-concepts, which are then added to the front of the goal queue. This demonstrates a hierarchical problem-solving ability.
5. **Idle Wandering and Boredom**: If the AI has no active goals, it enters an idle state. It traverses its memory graph by randomly choosing a connected node to shift its focus to. Each cycle of idle wandering increases its boredom level.
6. **Novelty Seeking (The Boredom Response)**: When boredom crosses its threshold (BOREDOM_THRESHOLD), the AI is motivated to find something new. It creates an "Expand knowledge" goal based on its current focus, forcing it out of the idle loop and back into a goal-oriented state.

4. Results and Discussion

The system's behavior was recorded in a detailed log file (hizawye_runtime.log). Analysis of these logs reveals both the success of the core design and the limitations of its components.

4.1 Successful Emergence of Strategic Failure

A key finding from the logs was the successful execution of the strategic failure mechanism. The logs show a clear sequence of events:

1. The AI attempts to "Deepen understanding of 'belief system'" multiple times.
2. Each attempt fails validation due to the LLM's confused, echo-like responses. pain increases with each failure.
3. A CRITICAL log entry appears: Pain threshold reached for 'belief system'. Initiating strategic failure.
4. The AI's goal changes to "Break down the concept: 'belief system'".
5. On the next cycle, it successfully queries the LLM for sub-concepts and receives ['self-evaluation', 'agreement', 'individual beliefs'].
6. It then correctly adds these as new nodes to its memory and sets them as its next goals.

This sequence demonstrates that the AI is not a simple automaton. It can recognize a failing strategy, alter its own goals in response, and adopt a more effective, hierarchical approach to problem-solving. This is a powerful emergent behavior that was not explicitly programmed but arose from the interaction of the simple rules governing its state.

4.2 The Challenge of the Reasoning Core

The primary source of failure was the brittleness of the tinyllama model. The logs are filled with instances where the LLM, when faced with a prompt requiring synthesis, would default to rephrasing its instructions. While our validation logic was improved to catch most of these, this highlights a key challenge: the control script's intelligence is handicapped by the limitations of its reasoning tool.

Furthermore, the project experienced a TypeError crash when the LLM, tasked with creating a list of strings, instead returned a list of lists. This necessitated a code update to make the AI's parsing logic more resilient and capable of handling malformed data from its reasoning core.

5. Challenges and Future Work

This research has successfully demonstrated a proof-of-concept but also highlighted areas for future development.

- **Upgrading the Reasoning Core**: The most significant improvement would be to swap tinyllama with a more powerful, instruction-following LLM. A larger model would be less prone to prompt confusion, allowing the AI to succeed more often and exhibit more complex reasoning.
- **Refining Internal States**: The current model of pain and boredom is simple. Future work could involve more nuanced states. For example, "pain" could be divided into "confusion pain" (from failed tasks) and "conflict pain" (from discovering contradictory beliefs in its memory graph).
- **Sensory Input**: The AI is currently a "brain in a vat." A future version could be given the ability to "read" external data (e.g., a Wikipedia article) as part of a research goal, allowing it to incorporate new, external knowledge into its memory graph.
- **Long-Term Memory and Forgetting**: The AI currently remembers everything forever. A more advanced model could include a "memory decay" mechanism, where nodes and connections that are not frequently visited become weaker over time, simulating the process of forgetting.

6. Conclusion

Hizawye AI successfully models a functional, simulated consciousness. By combining a transparent mind-state, a dynamic knowledge graph, and a simple set of rules governing its internal drivers (pain, boredom, curiosity), the system exhibits complex, emergent behaviors that were not explicitly programmed. It demonstrates goal-oriented focus, learns from failure by adapting its strategy, and is driven by boredom to seek novelty. While limited by its reasoning core, the project serves as a robust and extensible framework for exploring the computational underpinnings of consciousness and provides a clear path for future research into more sophisticated models of an artificial mind.