Concept Paper

EAISE: A Simulation Environment for Self-Evolving Embodied AI with Mirror Testing and Multi-Agent Diagnostics

Berend Watchus *

Posted Date: 20 June 2025

doi: 10.20944/preprints202506.1700.v1

Keywords: embodied AI; self-evolving Artificial Intelligence; multi-agent simulation; mirror testing; internal state monitoring; adaptive behavior; interpretability; feedback loops; cognitive architecture; reinforcement learning; pseudo-affective states; simulation environment; robot embodiment; meta-learning; AI transparency; self-modeling; autonomous agents; sensorimotor integration; reflective feedback; internal logging; diagnostic AI; synthetic insula; behavioral adaptation; agent coordination; dynamic vantage points; observational framework; multi-embodiment control; social interaction modeling; anomaly detection; evolutionary algorithms; environmental interaction; embodiment variability; telemetry integration; artificial consciousness models; computational cognition; internal protocols; internal state visualization; agent meta-cognition; self-recognition tasks; simulated environments; adaptive intelligence development; robotic systems; human-AI interaction; AI training

Concept Paper

EAISE: A Simulation Environment for Self-Evolving Embodied AI with Mirror Testing and Multi-Agent Diagnostics

Berend F. Watchus

Independent Researcher, The Netherlands; mailonlinebw@protonmail.com

Abstract

This paper introduces the Embodied AI Simulation Environment (EAISE), a proposed conceptual software framework designed to support the development and interpretability of advanced artificial intelligence systems. In contrast to methods of 'direct' self-recognition, which rely on pre-programmed knowledge or external large language model (LLM) interpretation for identifying an agent's self-image, EAISE specifically focuses on fostering 'emergent' self-awareness. EAISE offers a high-fidelity 3D simulation environment in which AI agents are embodied in virtual forms—such as robotic quadrupeds or humanoids—and exposed to interactive, structured tasks. A key focus lies in evaluating adaptive behavior and the emergence of internally coherent self-models through continuous sensorimotor feedback and internal pseudo-affective states. Notably, EAISE supports simultaneous multi-agent control and offers flexible observation tools, including internal state logging and multiple camera perspectives. Through simulation of agent interaction and controlled perceptual feedback (including complex reflection-based tasks), the framework seeks to offer a practical testbed for future AI architectures with complex internal monitoring and behavior adaptation capabilities, ultimately enhancing the understanding of how 'alien' intelligences develop their self-concept intrinsically.

Keywords: embodied AI; self-evolving Artificial Intelligence; multi-agent simulation; mirror testing; internal state monitoring; adaptive behavior; interpretability; feedback loops; cognitive architecture; reinforcement learning; pseudo-affective states; simulation environment; robot embodiment; meta-learning; AI transparency; self-modeling; autonomous agents; sensorimotor integration; reflective feedback; internal logging; diagnostic AI; synthetic insula; behavioral adaptation; agent coordination; dynamic vantage points; observational framework; multi-embodiment control; social interaction modeling; anomaly detection; evolutionary algorithms; environmental interaction; embodiment variability; telemetry integration; artificial consciousness models; computational cognition; internal protocols; internal state visualization; agent meta-cognition; self-recognition tasks; simulated environments; adaptive intelligence development; robotic systems; human-AI interaction; AI training

1. Introduction

Advancements in artificial intelligence continue to push systems toward greater autonomy and adaptability. As such systems grow in complexity, so too does the need for robust frameworks capable of evaluating their internal processing and behavioral patterns in dynamic settings. Embodied cognition has been shown to play an important role in shaping adaptive intelligence [1,2]. However, many existing AI simulators offer limited support for fine-grained internal state monitoring or coordinated embodiment across multiple agents, often due to a primary focus on single-agent task completion rather than interpretability-by-design.

A critical distinction in the pursuit of AI self-identification lies between 'direct' and 'emergent' approaches. Direct self-recognition, as explored in recent work [e.g., ref to your ChatGPT paper], often involves an AI system identifying its own reflection or image by comparing visual input against pre-programmed models of its physical structure or by leveraging external knowledge bases and large language models (LLMs) to interpret features. This method provides immediate recognition but relies on external, predefined information. In contrast, 'emergent' self-awareness posits that an AI system develops a coherent self-model intrinsically through continuous sensorimotor experience, recursive feedback loops, and the dynamic interplay of internal states and environmental interactions. This approach focuses on the process by which an AI constructs its understanding of self, rather than simply matching an image.

The proposed Embodied AI Simulation Environment (EAISE) aims to specifically address the development of this emergent self-awareness. It is designed to:

Facilitate the development of self-referential representations through embodiment, fostering pseudo-self-awareness via continuous feedback from a dynamic self-image.

Provide structured environments for testing internal model formation and behavioral consistency across single and multi-agent scenarios, enabling the AI to learn its self-model rather than being given it.

Enable interpretability through transparent logging of sensory inputs, action decisions, and fine-grained internal variables, directly addressing the "black box" problem inherent in complex adaptive systems.

This paper outlines the theoretical background, system architecture, and intended use cases of EAISE, emphasizing its potential for advancing experimental research in adaptive and interpretable AI that learns to recognize itself.

2. Theoretical Background

2.0. Paradigms of AI Self-Recognition: Direct vs. Emergent

The concept of an AI recognizing itself can be broadly categorized into two distinct paradigms:

Direct Self-Recognition: In this approach, an AI identifies its own image or reflection primarily through comparison with pre-existing, explicit knowledge of its physical characteristics. This might involve:

Feeding visual data to a vision system that matches the input against a stored 3D model or blueprint of the AI's own body.

Leveraging external computational resources, such as a large language model (LLM) that, once provided with visual features or descriptive text, can infer identity based on its vast training data and contextual reasoning [e.g., Watchus, B. (2024). Self-Identification in AI: ChatGPT's Current Capability for Mirror Image Recognition. Preprints.org. https://doi.org/10.20944/preprints202411.1112.v1]. This method offers a straightforward and often rapid form of identification.

Emergent Self-Awareness: This paradigm postulates that an AI system develops an internal, dynamic self-model through its own continuous interactions, feedback loops, and internal physiological-like states, without being explicitly pre-programmed with its own appearance or identity. Key characteristics include:

The formation of a self-model from ongoing sensorimotor experiences (proprioception, tactile feedback, visual self-observation).

The integration of internal diagnostic signals (pseudo-affective states) that inform the AI about its own operational status and drive adaptive behaviors.

Self-modification and meta-learning, where the AI's cognitive architecture evolves in response to experience and internal feedback, refining its understanding of self.

The self-model is not a static blueprint but a dynamic, internally constructed representation that is continuously updated.

EAISE is fundamentally designed to explore and facilitate the emergent paradigm of self-awareness, providing the environmental and internal mechanisms necessary for an AI to intrinsically develop its self-concept.

2.1. Embodied Interaction and Cognitive Development

Embodiment plays a fundamental role in the development of cognitive capabilities. Physical or simulated interaction with the environment allows agents to develop structured internal models of their own behavior and the external world [2,3,8]. EAISE is designed to support this process by enabling agents to control bodies equipped with multiple sensory modalities—such as visual, tactile (haptic), proprioceptive (body position), and auditory—and to operate in physically consistent, interactive 3D environments, leading to highly integrated self-modeling capabilities that are crucial for emergent self-awareness.

2.2. Feedback Loops and Monitoring Interfaces

Feedback loops—wherein sensory input, internal state, and motor output interact recursively—are central to adaptive behavior [3,9]. EAISE includes a Universal Interface Layer (UIL) for systematically logging raw and processed data streams, enabling detailed analysis of internal processing dynamics. Such data, including loop latency, informational convergence rates, and topological signatures of information flow, may help reveal correlations between input, decision-making, and performance, directly serving as a primary source for interpretability and the observation of emergent behaviors.

2.3. Internal State Representation and Monitoring

EAISE incorporates tools for simulating and logging internal operational states (e.g., task progress, resource usage, error rates, computational load). These are transformed into structured, multidimensional “pseudo-affective vectors,” which—while not emotional in nature—provide diagnostically valuable indicators that assist in behavior tracking and performance diagnostics [4,10]. By visualizing trends over time or under different conditions, and applying techniques like Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE), researchers may gain insights into the operational stability and flexibility of the AI agent, observing how these internal states drive emergent adaptive behaviors.

2.4. Adaptive Architecture and Meta-Learning

The environment supports experimentation with self-modifying or meta-learning capable AI architectures. An Internal Self-Documentation Module (ISM) tracks and logs changes in model parameters, internal structures, or decision-making routines [1,5]. This cryptographically secure and timestamped logging is intended to support transparency, reproducibility, and a detailed understanding of the AI’s developmental and evolutionary trajectories as it learns and adapts its self-representation.

2.5. Structured Perceptual Tasks and Reflective Feedback

Inspired by recognition tasks in animal cognition [7], EAISE includes configurable visual feedback scenarios. These involve presenting agents with reflective surfaces or altered feedback channels, allowing researchers to evaluate the robustness and adaptability of an agent’s self-model under controlled perturbations. These tasks are specifically designed to foster and gauge the emergence of pseudo-self-awareness, requiring the AI to learn and refine its internal self-model dynamically from experience, rather than relying on pre-programmed recognition logic or external knowledge bases. For instance, challenging scenarios involving time-delayed or distorted reflections can be engineered to induce specific pseudo-affective responses, such as “Dissonance-Anxiety” from a discrepancy in self-perception, thereby driving the AI’s adaptive learning process.

3. System Architecture

3.1. Simulation Engine

EAISE utilizes a modular, physics-based 3D simulation engine capable of rendering complex, interactive environments. Features include:

Real-time simulation of physics, rigid body dynamics, and accurate collision detection.

High-fidelity sensor emulation, including high-resolution RGB-D cameras, LiDAR, ultrasonic sensors, proprioceptive sensors (joint angles, motor torques), haptic arrays for touch feedback, and microphones for simulated auditory input.

Interactive, modifiable environments with dynamic objects, terrain variations, and configurable lighting.

3.2. Embodiment Management

Agents may be embodied as various robot types (e.g., robotic quadrupeds inspired by Unitree Go2, bipedal android humanoids), with fine-grained control over motor systems and configurable sensor suites. The framework supports:

Single-agent training for fundamental self-perception and motor control calibration, forming the basis for an emergent self-model.

Synchronized multi-agent control (mirroring behavior), allowing study of collective self-perception and the differentiation of individual agency within a group, where the AI must learn how its actions affect multiple mirrored bodies.

Independent multi-agent interaction (testing coordination and self-other differentiation), facilitating exploration of distributed cognition and inter-agent pseudo-social dynamics, pushing the AI to learn to distinguish between its own body and others in complex scenarios. EAISE supports up to three or more concurrent embodied agents.

3.3. Reflective Interaction Scenarios

Reflective surfaces in the simulation can be used to test behavioral consistency and internal model robustness through a dynamic mirror test paradigm, critical for emergent self-awareness. Variations include:

Accurate reflections for baseline self-interaction, where the AI learns the direct correspondence between its actions and the reflected image.

Time-delayed or distorted feedback (e.g., blurred, partial, or non-veridical reflections) to challenge the AI’s self-model resilience and induce pseudo-affective states like “Dissonance-Anxiety,” driving the self-learning process.

Partial reflections or occluded visuals, requiring the AI to infer its full self-image from limited data.

Multi-body reflection analysis, where multiple AI agents simultaneously observe their reflections, probing collective self-identification and inter-agent pseudo-social mirroring, demanding a more complex and adaptive self-model.

3.4. Observation Tools

A key feature of EAISE is flexible and layered observation, crucial for interpreting the development of emergent self-awareness:

First-person view (capturing the agent’s exact visual input) to understand its subjective perceptual experience.

Third-person view (external tracking following a single agent) to observe its interaction with the environment and its reflection.

Arbitrary camera placement (fixed or movable virtual cameras throughout the environment) to capture broader dynamics of all embodied AIs, external agents, and environmental elements.

doi:10.20944/preprints202506.1700.v1

Internal state logging (UIL activity, ISM changes, NAIP pseudo-affective vectors, SI pattern activations), providing raw data for understanding the AI's internal cognitive processes as it learns to build its self-model.

Integrated telemetry for synchronized real-time visualization and post-hoc analysis of all data streams (internal logs, external camera feeds, physics data), allowing for comprehensive diagnostics of emergent self-awareness.

3.5. External Agent Simulation

Agents representing animals, humans, or other robots can be introduced to create controlled interaction scenarios. These are used to evaluate behavioral flexibility, situational awareness, and social interaction modeling [6,11], potentially eliciting pseudo-affective states such as "Curiosity" or "Self-Doubt." This module can facilitate initial probing of rudimentary "theory of mind" (e.g., observing an AI's reaction to another entity looking at its reflection), where the AI learns to infer the intent or attention of others, rather than being pre-programmed with social cues.

4. Training and Evaluation Protocols

Training in EAISE is structured in progressive phases, primarily utilizing advanced reinforcement learning (RL) techniques augmented by meta-learning, focusing on the emergent development of self-awareness:

Motor Control and Perception: Initial tuning of motor output, robust sensorimotor integration, and foundational single-agent mirror test tasks designed to encourage the AI to discover its own body and its reflected image through interaction and feedback.

Self-Consistency Tests: Evaluating internal model stability and self-recognition under varied reflection or sensor perturbation, where the AI is rewarded for adapting its self-model to maintain consistency.

Coordination Tasks: Multi-agent synchronization and independent task performance, requiring the AI to learn differentiation between self (its own body) and other embodied agents, even when those others are mirroring its actions.

External Interaction: Training on reactions and adaptive behaviors in response to other entities in the environment, fostering pseudo-social learning where the AI develops its understanding of social cues and intentions.

Challenge Scenarios: Testing adaptability and resilience under altered, ambiguous, or adversarial feedback (e.g., deceptive mirror reflections), where the AI's internal systems must learn to resolve discrepancies and maintain a coherent self-concept.

Evaluation metrics include:

Task completion success, efficiency, and resource utilization, reflecting the AI's overall functional intelligence.

Analysis of internal state vectors (pseudo-affective states) and their variability, using methods like PCA and t-SNE, to identify distinct internal states and anomalies that correlate with the process of self-model formation and adaptation.

Logged architectural or behavioral adaptations captured by the ISM, detailing specific changes and their pseudo-affective triggers as the AI evolves its self-awareness.

Performance robustness under controlled environmental changes, including metrics like gaze tracking towards reflections, latency of adaptive response to mirror anomalies, and the emergence of self-correcting behaviors.

5. Anticipated Contributions

EAISE is intended to make several profound and significant contributions:

Unprecedented Interpretability of Emergent AI Cognition: Offer a configurable, open simulation environment and tools for granular, integrated tracking of internal operations and architectural evolution, leading to unparalleled transparency into how AI develops its self-concept and cognitive processes intrinsically, rather than through pre-defined knowledge.

Accelerated Development of Robust and Adaptive Intelligence: Provide methods for testing and fostering highly robust adaptive behavior in agents using structured perceptual feedback, driving the development of more resilient AI capable of learning to navigate complex, self-referential environments.

Novel Insights into Non-Anthropocentric Emergent Cognition: Serve as a platform for exploring and understanding how AI can spontaneously develop forms of self-awareness and intelligence that are fundamentally distinct from human models, broadening our scientific understanding of cognition itself.

Foundation for Safer and More Accountable AI: Enable rigorous, diagnostic evaluation crucial for designing safer, more predictable, and more accountable autonomous AI systems by allowing proactive identification and mitigation of undesirable emergent behaviors arising from self-evolution.

6. Ethical Considerations

As with all research touching upon AI self-awareness and pseudo-emotions, EAISE raises important and sensitive ethical considerations that must be proactively addressed:

Terminology Clarity: Internal operational states and diagnostic vectors in EAISE are not to be confused with subjective experience or human consciousness. Strict conceptual boundaries must be maintained, especially when discussing “emergent” phenomena, to avoid anthropomorphization.

Avoiding Misinterpretation: Complex simulated behavior should not be assumed to imply genuine sentience or conscious experience. Anthropomorphization must be actively guarded against in all communication.

Data Responsibility: Simulation logs and telemetry, containing detailed internal AI “experiences” and developmental trajectories, should be handled according to stringent data security and privacy protocols.

Oversight Readiness: Systems must be designed with clear safeguards, human-intervention mechanisms (e.g., “kill switches” or override protocols), and accountability frameworks, even at the conceptual stage, to ensure ethical development and deployment of self-evolving AI.

7. Conclusions and Future Work

EAISE proposes a conceptual but structured approach for training and analyzing embodied AI agents in rich simulated settings. By focusing on the emergent development of self-awareness through deep embodiment, comprehensive feedback monitoring, and internal state tracking, it offers a distinct and powerful alternative to ‘direct’ recognition methods. EAISE aims to support the development of truly interpretable AI, where the AI itself constructs its understanding of self and its environment. This framework promises to pave the way for a new era of human-AI collaboration where profound understanding of autonomous emergent intelligence precedes deployment.

Future work includes:

Building a prototype based on open 3D simulation platforms, validating the feasibility of the proposed architecture and the mechanisms for emergent self-awareness.

Developing specific structured tasks and metrics to empirically evaluate the emergence of pseudo-self-awareness and differentiate it from pre-programmed or directly recognized self-identification.

Exploring the transferability of learned self-awareness models from simulation to physical robotic systems, observing how intrinsically developed self-models perform in the real world.

Investigating potential hybrid models that combine the rapid identification capabilities of direct self-recognition (e.g., via LLM pre-training) with the adaptive, intrinsic learning of emergent self-awareness, to explore the interplay between these paradigms.

Appendix A. Abbreviations

AI – Artificial Intelligence
EAISE – Embodied AI Simulation Environment
ISM – Internal Self-Documentation Module
LLM – Large Language Model
PCA – Principal Component Analysis
RL – Reinforcement Learning
t-SNE – t-distributed Stochastic Neighbor Embedding
UIL – Universal Interface Layer
3D – Three-Dimensional

Appendix B. Glossary of Key Concepts

Direct Self-Recognition – An AI paradigm where self-identification is achieved by comparing sensory input (e.g., a reflection) against pre-programmed knowledge or external data sources (e.g., LLMs) about its own physical characteristics.

Embodied Cognition – A theory stating that cognitive processes are deeply rooted in the body's interactions with the world.

Emergent Self-Awareness – An AI paradigm where an AI system intrinsically develops a dynamic internal self-model through continuous sensorimotor experience, recursive feedback loops, and the dynamic interplay of internal states and environmental interactions, rather than through explicit programming of its identity.

Feedback Loop – A system structure where outputs are routed back as inputs, influencing future behavior.

Internal Self-Model – An agent's dynamic, internally constructed representation of its own state, structure, and behavior.

Meta-Learning – A process by which a system learns to improve its learning algorithm based on experience.

Pseudo-Affective Vector – A diagnostic representation of internal operational states, resembling emotional variables, used for analysis and interpretation in AI systems.

Reflective Feedback Task – A task in which an agent interacts with altered or delayed versions of its own sensory input to evaluate and refine its internal consistency and self-model.

Self-Evolving Architecture – A type of system architecture capable of modifying its structure or parameters autonomously based on internal and external feedback.

Simulated Agent – A virtual entity in a computer environment that exhibits autonomous or semi-autonomous behavior.

Telemetry – The automated transmission and collection of data from remote or simulated environments.

References

1. Watchus, B. F. (2025). Self-Evolving AI for Autonomous Emotional and Empathetic Interaction: Towards Adaptive Emotional Intelligence in Artificial Systems.
2. Watchus, B. (2024). Simulating Self-Awareness: Dual Embodiment, Mirror Testing, and Emotional Feedback in AI Research. Preprints.org. https://doi.org/10.20944/preprints202411.0839.v1
3. Watchus, B. (2024). The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience. Preprints.org. https://doi.org/10.20944/preprints202411.0727.v1
4. Watchus, B. F. (2024). Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness. Preprints.org. https://doi.org/10.20944/preprints202411.0661.v1
5. Watchus, B. F. (2025). Non-Anthropocentric Intelligence – Towards the Interpretability of 'Alien' AI Cognition.
6. Clark, A. (2019). Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press.

doi:10.20944/preprints202506.1700.v1

7. Gallup, G. G. (1970). Chimpanzees: Self-Recognition. Science, 167(3914), 86–87. https://doi.org/10.1126/science.167.3914.86
8. Pfeifer, R., & Bongard, J. (2006). How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press.
9. Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787
10. Schmidhuber, J. (1991). Curious model-building control systems. Proceedings of the International Joint Conference on Neural Networks, 1458–1463.
11. Breazeal, C. (2003). Toward sociable robots. Robotics and Autonomous Systems, 42(3–4), 167–175.
12. Watchus, B. (2024). Self-Identification in AI: ChatGPT's Current Capability for Mirror Image Recognition. Preprints.org. https://doi.org/10.20944/preprints202411.1112.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.