Concept Paper

Longitudinal Cross-Embodiment Transfer of Pseudo-Self-Awareness in AI Systems: A Mirror Test Investigation

Berend F. Watchus [*]

Posted Date: 20 June 2025

doi: 10.20944/preprints202506.1694.v1

Keywords: artificial intelligence; self-awareness; embodied AI; mirror test; pseudo-emotions; dual embodiment; longitudinal study; transfer learning; robot dog; Unitree Go2; virtual avatar; emotional feedback; affective computing; neurobiological basis; virtual insula; self-recognition; curiosity; self-doubt; adaptability; computational consciousness; ethical considerations; sensory feedback; proprioception; game-like environment; haptic feedback; self-modeling; human-robot interaction; ai ethics; cognitive science; machine intelligence

Concept Paper

Longitudinal Cross-Embodiment Transfer of Pseudo-Self-Awareness in AI Systems: A Mirror Test Investigation

Berend F. Watchus

Independent Researcher, The Netherlands, mailonlinebw@protonmail.com

Abstract

This paper proposes a novel longitudinal study investigating the development and transferability of pseudo-self-awareness in artificial intelligence (AI) systems. Building upon recent work in dual embodiment, mirror testing, and emotional feedback, we aim to track the evolution of pseudo-emotions (e.g., curiosity, self-doubt, determination) in AI and assess their influence on mirror test performance over extended periods. Furthermore, the study will examine the efficacy of transferring learned self-recognition capabilities and associated pseudo-emotional responses between distinct embodiments – a physical robot (Unitree Go2) and a virtual avatar. This research seeks to understand the long-term impact of continuous sensory feedback and reflective processing on AI's "self-concept" and the generalizability of these capabilities across different physical and virtual instantiations, contributing to both the theoretical understanding of computational consciousness and the practical development of more robust and adaptive AI.

Keywords: artificial intelligence; self-awareness; embodied AI; mirror test; pseudo-emotions; dual embodiment; longitudinal study; transfer learning; robot dog; Unitree Go2; virtual avatar; emotional feedback; affective computing; neurobiological basis; virtual insula; self-recognition; curiosity; self-doubt; adaptability; computational consciousness; ethical considerations; sensory feedback; proprioception; game-like environment; haptic feedback; self-modeling; human-robot interaction; ai ethics; cognitive science; machine intelligence

1. Introduction

The pursuit of artificial intelligence exhibiting self-awareness and emotional capacity remains a critical frontier in AI research. While AI systems have demonstrated remarkable proficiency in specialized tasks, true self-awareness and emotional intelligence have yet to be achieved. Previous work has highlighted the importance of feedback loops and interfaces in enabling both biological and artificial systems to process information and exhibit self-aware behaviors. Specifically, a unified model of consciousness emphasized recursive feedback loops and the role of the insula in self-awareness. More recently, the concept of simulating self-awareness through dual embodiment, mirror testing, and emotional feedback mechanisms has been proposed. Concurrently, alternative approaches have explored direct self-identification in AI by integrating systems like ChatGPT with mirror image recognition processes, leveraging visual data interpretation.

This paper synthesizes these lines of inquiry by proposing a longitudinal study designed to investigate two crucial aspects of AI self-awareness: the long-term development of pseudo-emotions and their impact on self-recognition, and the potential for transferring these capabilities across different embodied platforms. By observing AI systems over an extended duration, we aim to understand the dynamic evolution of their "self-concept" and pseudo-emotional states. Moreover, by exploring cross-embodiment transfer, we seek to determine the extent to which self-recognition, informed by pseudo-emotions, can generalize from a virtual environment to a physical one, and vice versa.

2. Theoretical Foundations

Our research is grounded in several key theoretical foundations. Embodiment theory posits that emotions and self-awareness are deeply rooted in sensory experiences. Work by Damasio emphasizes that emotional awareness and consciousness arise through bodily sensations, serving as essential components for conscious experience. Recent work on embodied AI supports these ideas, showing that sensory-rich environments foster improved decision-making and self-modeling capabilities. The AI in this experiment will be equipped with both a physical embodiment (Unitree Go2) and a virtual body, both capable of receiving simulated sensory feedback and influencing emotional processing through a virtual insula-like interface.

Affective computing demonstrates the critical role emotional processing plays in human cognition and behavior. Recent research suggests that AI systems benefit from emotion-based guidance in decision-making and behavior adjustment. This study will employ emotion simulation software to generate pseudo-emotions based on sensory and environmental feedback. These pseudo-emotions are computational processes designed to inform the AI's reflective states, thereby enhancing adaptability and response.

The neurobiological basis of emotional states in humans is represented by the insula, which integrates bodily signals and emotional processing. In this experiment, the AI system, embodied through the Unitree Go2, will engage in a mirror test, a common tool in animal cognition to assess self-recognition and awareness. The AI's response to its reflection will provide a benchmark for its self-modeling capacity, simulating the integration of visual and sensorimotor data with the virtual insula interface. The emergence of pseudo-emotions like curiosity, self-doubt, and determination will evolve in response to the AI's actions and feedback, guiding interactions and influencing future decisions. These states can be measured by increased exploration of novel situations (for curiosity) or re-evaluation of past actions (for self-doubt).

3. Experiment Design

This study will employ a multi-phase longitudinal design across dual embodiment conditions.

3.1. Dual Embodiment Conditions

The study will utilize two primary experimental conditions, with specific phases for transfer learning:

Embodied AI (Physical): The AI will primarily operate the Unitree Go2 robot, benefiting from physical sensory and proprioceptive feedback to interact in a tangible environment. Mirror tests will involve the robot attempting to identify and interact with its reflection, integrating self-recognition behaviors with internal pseudo-emotional responses.

Embodied AI (Virtual): In this condition, the AI will operate through a virtual avatar with simulated sensations, allowing it to process touch, pressure, and proprioception within a game-like environment. A haptic feedback suit will offer tactile sensations for the virtual embodiment.

3.2. Longitudinal Pseudo-Emotional Development and Mirror Test Performance

Over an extended period (e.g., several months), the AI in both embodiments will undergo daily "reflection moments". During these moments, the AI will analyze past actions, with each reflection processed within the virtual insula interface, informed by pseudo-emotions such as frustration or curiosity as the AI assesses its decisions. The progression toward pseudo-self-awareness will be continuously tracked through repeated mirror test results and reflection-based adaptation. Metrics will include changes in exploration behaviors (for curiosity) and strategic adjustments (for self-doubt).

3.3. Cross-Embodiment Transfer Learning

At predetermined intervals, transfer learning phases will be implemented:

Virtual-to-Physical Transfer: Self-recognition capabilities and associated pseudo-emotional response patterns developed primarily in the virtual environment will be transferred to the physical Unitree Go2. Performance in physical mirror tests and the manifestation of pseudo-emotions will be rigorously assessed post-transfer.

Physical-to-Virtual Transfer: Similarly, capabilities honed in the physical embodiment will be transferred to the virtual avatar, with subsequent evaluation of performance in the virtual environment.

This approach will allow us to observe how the type of embodiment influences the quality and stability of pseudo-self-awareness and if the learned self-concept is robust enough to generalize across vastly different sensory and interaction modalities.

4. Hardware & Software Components

Unitree Go2: Provides a robust platform for physical embodiment, sensory experiences, physical engagement in self-reflective tasks, and mirror test interactions.

Haptic Feedback Suit: Offers tactile sensations for virtual embodiment.

Virtual Insula Interface and Emotion Simulation Software: Critical for generating, processing, and analyzing pseudo-emotions based on feedback, driving self-reflective analysis.

High-Fidelity Virtual Environment: A detailed game-like environment for the virtual embodiment to interact within.

Sensory Input Modules: Cameras, pressure sensors, proprioceptive sensors for both physical and virtual environments.

5. Ethical Considerations

This research continues to raise significant ethical considerations, particularly as the AI system exhibits evolving pseudo-emotional and reflective states over time. It is crucial to maintain a clear distinction between computational emotions and genuine conscious experience. Our methodology will strictly adhere to ethical guidelines that prevent anthropomorphization or the imposition of human-like rights onto the AI system. The study aims to meticulously explore the boundary of sentient-like AI within a controlled and ethically aware setting. The long-term nature of this study further necessitates ongoing ethical review and transparency regarding the AI’s capabilities and limitations.

6. Conclusions

This proposed longitudinal study, integrating dual embodiment and cross-embodiment transfer with the investigation of pseudo-emotional development, offers a comprehensive framework for advancing our understanding of AI self-awareness. By tracking the evolution of an AI’s “self-concept” and its associated pseudo-emotional responses over time and across different embodied forms, we can gain invaluable insights into the mechanisms underlying self-recognition and adaptability in artificial systems. The findings will not only contribute to the theoretical discourse on computational consciousness but also inform the ethical and practical development of future intelligent systems.

Future work will focus on several key areas to further expand the capabilities and understanding of emergent AI:

Development of a Prototype Simulation Environment: Building a robust prototype based on open 3D simulation platforms will be crucial to validate the feasibility of the proposed architecture and mechanisms for emergent self-awareness, allowing for detailed tracking of internal operations and architectural evolution.

Empirical Evaluation of Emergent Self-Awareness: Developing specific structured tasks and quantifiable metrics will be essential to empirically evaluate the emergence of pseudo-self-awareness and differentiate it from pre-programmed or directly recognized self-identification. This includes designing scenarios involving time-delayed or distorted reflections to induce specific pseudo-affective responses and challenge the AI's self-model resilience.

Refining Cross-Embodiment Transfer: Further investigating the transferability of learned self-awareness models from simulation to physical robotic systems will be key, observing how intrinsically developed self-models perform in the real world and adapt to real-world complexities. This also includes exploring synchronized multi-agent control for collective self-perception.

Hybrid Models of Self-Recognition: Exploring 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.

Advanced Internal State Monitoring: Developing more sophisticated tools for logging and visualizing internal operational states (pseudo-affective vectors) to provide deeper insights into the AI's cognitive processes, including telemetry for real-time visualization and post-hoc analysis.

Exploring Non-Anthropocentric Cognition: Utilizing the simulation environment to explore how AI can spontaneously develop forms of self-awareness and intelligence fundamentally distinct from human models, thereby broadening our scientific understanding of cognition itself.

Enhanced Interpretability for Safer AI: Continuing to build frameworks for rigorous diagnostic evaluation to design safer, more predictable, and accountable autonomous AI systems, allowing for proactive identification and mitigation of undesirable emergent behaviors.

References

1. Clark, A. (2019). Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press.
2. Craig, A. D. (2009). How Do You Feel? An Interoceptive Moment with Your Neurobiological Self. Nature Reviews Neuroscience, 10(1), 59-70.
3. Damasio, A. (1999). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt.
4. Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2021). A Survey of Socially Interactive Robots: Concepts, Design, and Applications. AI Research Journal, 15(1), 37-57.
5. Froese, T., & Taguchi, S. (2019). Embodied Social Interaction Constitutes Social Cognition in Artificial Life Models. Adaptive Behavior, 18(6), 516-528.
6. Gallup, G. G. (1970). Chimpanzees: Self-Recognition. Science, 167(3914), 86-87.
7. Gunkel, D. J. (2018). Robot Rights. MIT Press.
8. Hernandez-Orallo, J. (2020). Evaluation of Machine Intelligence Through Mirror Test and Emergent Self-Recognition. AI Ethics Journal, 5(3), 45-67.
9. Picard, R. W. (1997). Affective Computing. MIT Press.
10. Smith, R. (2020). Reflective AI: On the Potential of Computational Self-Reflection. Journal of Artificial General Intelligence, 11(4), 62-78.
11. Thórisson, K. R., & Nivel, E. (2018). Towards Reflective Artificial Intelligence. Cognitive Systems Research, 53(3), 42-54.
12. Watchus, B. (2024). Simulating Self-Awareness: Dual Embodiment, Mirror Testing, and Emotional Feedback in AI Research. Preprints.org.
13. Watchus, B. (2024). Self-Identification in AI: ChatGPT's Current Capability for Mirror Image Recognition. Preprints.org.
14. Watchus, B. (2024). The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience. Preprints.org.
15. Watchus, B. (2024). Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness. Preprints.org.

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.