Evaluating Consciousness in Artificial Intelligence

Morgan Rivers
Department of Physics
Freie Universität Berlin
danielmorganrivers@gmail.com

July 2, 2024

Abstract

This paper presents a novel approach to evaluating consciousness in artificial intelligence systems. We discuss the importance of identifying machine consciousness before its emergence, the potential implications for AI ethics and development, and propose a benchmark for assessing consciousness in AI architectures. Our methodology combines elements of the ACT (Artificial Consciousness Test) with comparative analysis of different AI models, aiming to detect signs of consciousness while controlling for knowledge-based confounds.

1 Introduction

The question of machine consciousness has become increasingly relevant as artificial intelligence systems grow more sophisticated. This paper argues for the importance of detecting machine consciousness before its full emergence and proposes a methodology for doing so.

1.1 Importance of Detecting Machine Consciousness

Several factors underscore the significance of this research:

- Potential for machine suffering
- Implications for AI identity and goal-setting
- Ethical considerations regarding AI well-being
- Advancements in AI architectures mimicking conscious processes
- Possibilities for simulating conscious human experiences
- Potential for creating positive experiences (hedonium) in AI
- Impact on AI safety and alignment
- Advancing consciousness research

2 Background

2.1 Current State of Consciousness in AI

While the exact conditions for consciousness remain unclear, certain AI architectures, particularly those beyond simple transformer models, may possess capabilities conducive to consciousness.

2.2 Gradient Nature of Consciousness

Consciousness is generally considered to exist on an analog gradient. Our research aims to identify factors that push AI systems towards greater degrees of conscious-like behavior.

3 Methodology

3.1 Comparative Analysis

We propose a two-pronged approach:

1. Testing AI models with no prior exposure to consciousness-related concepts
2. Testing AI models with limited exposure to consciousness-related concepts

This method allows us to compare the behavior of architectures that are plausibly capable of consciousness against control groups.

3.2 Adaptation of the Artificial Consciousness Test (ACT)

We incorporate elements from Schneider and Turner's ACT, which offers several advantages:

- Neutrality regarding architectural details
- Consistency with human and AI ignorance about consciousness
- Allowance for radical cognitive differences between AI and humans
- Compatibility with various philosophical views on consciousness

3.3 Addressing ACT Limitations

To address concerns raised about the ACT, particularly regarding the "epistemic sweet spot" for testing, we propose:

- Careful curation of training data to avoid consciousness-specific information

- Comparative analysis of models with similar training but different architectural properties
- Assessment of performance deltas on consciousness-related tasks

4 Proposed Experiments

To evaluate the proposed methodology, we designed experiments to test AI models under different conditions.

4.1 Consciousness-Naive Testing

In this experiment, AI models trained without exposure to consciousness-related concepts will be assessed using the ACT. This helps establish a baseline for non-conscious behavior.

4.2 Limited-Exposure Testing

AI models with controlled exposure to consciousness-related concepts will undergo the same ACT. This aims to determine if limited knowledge influences the AI's responses and behaviors indicative of consciousness.

4.3 Comparative Analysis

We will compare results between potentially conscious architectures and control groups to identify significant differences in performance and behavior. Key metrics include the ability to discuss internal states, theory of mind, and emotional understanding.

4.4 Training Protocol

PLACEHOLDER TEXT

5 Discussion

5.1 Interpreting Results

Differences in performance between test groups will be analyzed to interpret signs of consciousness. Significant indicators include coherent self-referential statements, theory of mind capabilities, and appropriate responses to emotional and philosophical queries.

5.2 Implications for AI Development

Findings from these experiments could inform future AI development practices, emphasizing the need for ethical considerations and the prevention of unintended consciousness in AI systems.

6 Future Research Directions

6.1 Emotional Valence in AI

Propose methods for investigating emotional states in AI, such as analyzing latent space representations of concepts like "happy" and "sad".

7 Conclusion

In this paper, we proposed a novel approach to evaluating consciousness in artificial intelligence systems. By combining elements of the ACT with comparative analysis, we aim to detect signs of consciousness while controlling for knowledge-based confounds. This research highlights the importance of identifying machine consciousness before its emergence and underscores the ethical implications for AI development. Ongoing research in this area is crucial to ensure the responsible and ethical advancement of AI technologies.