A Benchmark for Consciousness in Large Language Models

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

July 3, 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), which aims to detect robust indicators of consciousness while controlling for knowledge-based confounds. A simple training scheme to induce the emergence of consciousness in small open source language models is also introduced.

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.

While the exact conditions for consciousness remain unclear, certain AI architectures, particularly those beyond simple transformer models, may possess capabilities conducive to consciousness. A test has been proposed by Schneider et al. (Turner and Schneider, 2018) and expanded on by Perez et al. (Perez and Long, 2023) and Long (**?**) that attempts to determine whether a language-enabled AI agent is conscious. Specifically, Perez and Long recommend the training for introspection of language models in order for consciousness to arise. While such approaches are interesting, the task remains to actually build a benchmark and training scheme for LLM's such that they can be tested at all, before diving into tweaks on a method of training involving introspection. The key practical questions remain unanswered in the literature: 1. How to introduce language models to language, without or with minimal examples generated

from conscious beings (e.g. humans) to ensure they are not simply mimicking consciousness, rather than emergently exhibiting conscious behavior? 2. How to ensure such a training scheme can enhance language ability without language examples? 3. What key aspects of consciousness in a training scheme can both be learned reliably, and be reasonably expected to produce conscious behavior?

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

1.2 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.

1.3 Proposed method of achieving benchmark

Social animals that are widely regarded as conscious share the property that they have developed a theory of mind, and exhibit social behavior. Such animals over the aeons have emergently developed consciousness without other conscious beings training them on every behavior. We therefore have empirical evidence that selective pressures under the condition of socialization are therefore key to allow for consciousness to emerge.

We are left with the observation that in order for consciousness to emerge, language itself must also emerge in the process without continued examples of natural language, as occurred in the animal kingdom through evolutionary selective pressures.

We propose to first initialize an untrained transformer model of modest size (100 million parameters) and teach it rudimentary logic and language with rules generated from ProofWriter (Tafjord et al., 2021). By utilizing a small number of rules, it is possible to teach the model to both follow logical steps, and to use certain keywords in an appropriate way in the game of Diplomacy.

Next, we introduce the model to the board game diplomacy as was implemented in (Meta Fundamental AI Research Diplomacy Team et al., 2022) the following training scheme to meet the benchmark described above. By playing against itself in a strategic game, it will utilize the reasoning and ability to use words as meaningful tokens. Full-press Diplomacy is a game where effective cooperation, planning, and theory of mind are critical to effective gameplay. The secondary benefit to using the game is it mimics evolutionary environments of conscious animals, where cooperation, strategy, and theory of mind are critical to achieving reward. Finally, full-press diplomacy has been shown in (Meta Fundamental AI Research Diplomacy Team et al., 2022) to increase game performance over time with only self-play.

2 Methodology

2.1 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.

2.2 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

2.3 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

3 Proposed Experiments

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

3.1 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.

3.2 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.

3.3 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.

3.4 Training Protocol

PLACEHOLDER TEXT

4 Discussion

4.1 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 Future Research Directions

5.1 Emotional Valence in AI

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

6 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.

References

Turner, E., & Schneider, S. (2018, January). Testing for synthetic consciousness: The ACT, the chip test, the unintegrated chip test, and the extended chip test. In CEUR Workshop Proceedings (Vol. 2287). CEUR-WS.

Perez, E., & Long, R. (2023). Towards Evaluating AI Systems for Moral Status Using Self-Reports. *arXiv preprint arXiv:2311.08576*. Retrieved from `https://arxiv.org/abs/2311.08576`

Tafjord, Ø., Mishra, B. D., & Clark, P. (2021). ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. *arXiv preprint arXiv:2012.13048*. Retrieved from `https://arxiv.org/abs/2012.13048`

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., ... & Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067-1074.