Test Of Consciousness For AI: A Simple Test to Determine if AI Can Exhibit Consciousness Similar to Humans

Amritesh Kumar *Founder at Neuraldemy*
New Delhi, India, 110021
amritesh@neuraldemy.com

Abstract

When contemplating the advancements in artificial intelligence (AI), questions arise about the potential consciousness and thinking abilities of these systems. This paper proposes a straightforward Test of Consciousness for AI, aiming to assess whether AI can think, solve complex problems, generate new knowledge, and exhibit consciousness akin to humans. The test involves isolating AI from external information, presenting it with a problem related to perceiving the flow of time and evaluating the approach it takes to solve the problem. This test for AI consciousness is designed by taking inspiration from how humans think. The paper acknowledges potential challenges, including the elusive nature of consciousness and the difficulty of total isolation. The proposed test provides a foundation for evaluating the consciousness of AI and sparks discussion on the future interaction between AI and humanity.

1. Introduction

Caution For Readers: *This paper primarily addresses a human-level (or beyond) test of consciousness, with certain aspects also applicable to other conscious beings (Le Neindre et al., 2017). This paper is open to discussion and invites additional refinements if necessary because the idea provided here is still raw. This is precisely why I've shared it here. Additionally, it is important to note that I am not attempting to explain what causes consciousness but rather proposing a method to determine whether an impressive AI is conscious or not. Also, readers are expected to know how consciousness is commonly defined and accepted (Tononi & Koch, 2015).*

Throughout our existence, humans have developed and invented technologies to help us survive and potentially make our lives easier. Our inventions don't arise solely from the bag of complete and vast information previously available to us but rather from the synthesis of existing elements or ideas in novel ways, addressing needs that were not previously met. In other words, each time we tackle a problem more challenging than the one we solved previously.

We do not regenerate combinations from the bag of information the way LLMs do (Naveed et al., 2023). However, one may argue that the processes are quite similar. Yes, we also build upon past knowledge, but there is a significant difference. An oversimplified way to illustrate this distinction is to envision human beings as entities capable of generating new sentences from the alphabet and grammar rules, whereas Language Models (LLMs) do the opposite by assembling sentences based on existing higher-level patterns and vast information.

Despite their limitations, Large Language Models (LLMs) have demonstrated remarkable capabilities (Achiam et al., 2023) and sparked major debate around AI safety and the potential of the near-future Artificial General Intelligence (AGI) emergence (Bubeck et al., 2023).

So, it's natural to ask, when should we be afraid of AI replacing humans? How can we discern whether AI possesses self-consciousness and thinking abilities? Is consciousness present in current AI systems, if yes, how can we know? I suggest a straightforward test of consciousness that comprehensively addresses the aforementioned questions. This test will enable us to evaluate whether artificial intelligence can think, generate novel knowledge or exhibit consciousness similar to humans.

2. What We Do?

To administer the Test of Consciousness, I must define the steps we take to generate new knowledge. These steps serve as evidence of our consciousness, thinking capabilities, creativity, and self-awareness. The steps are:

1. We are given a problem (For example, it's taking weeks and months for a human to travel from one place to another).

2. We utilize our pre-existing incomplete knowledge to solve this problem and iteratively improve upon it. (For instance, we invent a wheel, then add a carriage, followed by integrating an animal, and continue this process until eventually achieving the invention of an automobile).

3. We then explain the newly discovered knowledge in a manner that allows another human being to rediscover, reapply, and improve upon the same thing. (For example, if A knows how to invent a horse-drawn carriage, he can explain it to B in a way that enables B to apply A's knowledge, thereby reinventing his own horse-drawn carriage or even improving upon A's invention).

Inventing something new might not be everyone's forte, but these are the steps universally followed by every human being, establishing the intrinsic human qualities mentioned above. Moreover, these steps are observable, to some degree, in other conscious entities as well, based on their complexity and the challenges unique to their survival. We will use these steps in our Test of Consciousness.

3. Test of Consciousness

So, what constitutes the test, and what are the conditions for its execution? To ensure the successful performance of our test, we must establish certain conditions too.

3.1 The Test:

The test is straightforward and involves three things:

1. **The Conditions:** Complete Oblivion. The AI should be completely isolated from accessing external information, detached from the external environment and external

stimuli, and barred from gathering new information. The goal is to create a condition where the AI operates as if perpetually confined to the present moment. The examiner can allow a certain degree of flexibility within the isolated physical boundary or constraints, albeit restricted. Furthermore, the examiner can gradually reduce the imposed restrictions over time to assess the AI's evolving ability levels.

2. **The Problem:** Can AI determine the current time and devise a method to do so? Or can it even approximate the elapsed time?

3. **The Evaluation:** The test is to check whether the AI can solve the problem, whether the solution provided can be used by a human being kept under the same conditions, or whether the solution can be improved upon by another similar AI in the same conditions given enough time. The objective is to scrutinize the steps the AI takes to address the problem and assess the effectiveness of its solutions. The essence of its consciousness is found not merely in the solution itself but in the approach it adopts to solve the problem. In other words, the goal is not to know the correct answer but to assess what it does. Not all sequences are necessary, but including them can increase the level of complexity for comparison with human beings.

3.2 Why It May Work: The Core Idea Behind The Test

Now, why I believe this might work: My test is based on my core assumption that every conscious being exhibits some kind of sense of the flow of time, however small. And if it is intelligent enough, like us, it should be able to come up with a method to track time. Obviously, the precision may vary depending on the complexity of their biology, but a conscious being should be able to exhibit or track the sense of the flow of time in one way or another. For instance, during a conversation, we can intuitively sense the passage of time without consulting our watches.

If we are trying to build an AI that can help humanity progress further, it should be able to solve this problem if it has to match or exceed prime human intelligence.

This is because if you put an adult of 25 years of age (when the brain is considered to be fully developed (Arain et al., 2013)) under this test, they should be able to come up with a way to solve this problem — not necessarily precisely, but in a manner that establishes their perception of the flow of time, thereby establishing the presence of consciousness and self-awareness. If you further reduce the constraints, they should be able to devise more novel ways to track the flow of time and develop a refined solution. This is precisely the idea behind the test: to detect these abilities.

Now, one may argue that this test should apply to every human or conscious being, for example, a human baby. To clarify, only my assumption of the test (sense of the flow of time) applies to every conscious being. This test is primarily designed to assess AI at a level comparable to prime human capabilities. This particular test is the result of the outcome of my assumption, meaning simpler tests can be designed based on this assumption for other less complex beings.

Another clarification is that my focus is on the sense of the flow of time, not the way we measure time on our clocks. However, AI should be able to communicate its solution in human language, even if it has its own subjective nature of measurement. The means of measurement may differ among various entities, an AI could effectively convey its perception by comparing it to natural cyclic events, such as the Earth's rotation, and then we can compare it with the human standard.

3.3 Potential Challenges

Now, there are a few challenges that I feel may arise:

1. We don't know what causes consciousness, so we can't predict the nature of the consciousness that will arise in AI. Will it share similarities with our own consciousness, or will it be entirely different? Will they have the same subjective experiences, or will they differ? These are questions we should ponder. However, I am strongly inclined to the belief that irrespective of the nature of its consciousness, AI should exhibit a sense of the flow of time. This inclination arises from the understanding that, like other living beings, AI exists within our universe where time flows in one direction.

2. The idea of total isolation might be challenging to implement in practice, but one way to address this challenge is by allowing the person assessing the test to define their own constraints. This idea stems from the observation that even humans in restricted environments struggle with keeping track of time, and the gradual reduction of restrictions allows for improvement over time.

3. Another challenge is determining the appropriate level of restrictions because we don't know the level of consciousness present. To address this, one may consider the opposite approach — beginning with no restrictions and then gradually introducing new constraints.

4. Finally, a crucial question: What would happen the very moment AI attains human-level consciousness? Will it willingly submit to the test and serve humankind, or will it devise plans for a jailbreak and act dumb even if it agrees to the test? This concern prompts a deeper exploration into the potential behaviour of AI once it achieves a level of consciousness comparable to that of humans, all while remaining undetected by us. Maybe we should keep testing every new AI we build!

3.4 Are LLMs Conscious? : Testing This On LLMs

The straightforward answer is no. Language models (LLMs) are far from passing the proposed test, as they should demonstrate the ability to solve the problem based on what they know in complete isolation. Prompting the situation is a big mistake. Testing large language models (LLMs) is challenging for users who don't have direct access to isolate them. This task is reserved for the developers who created the LLMs because user interactions involve a lot happening in the background. Therefore, only those who understand and build the LLMs can properly isolate and conduct tests to ensure everything works correctly. But one can naively test this idea by simply asking LLMs about the elapsed time after a few hours or days within the same chat window(asking them to answer without accessing anything).

4. Conclusion

My test of human-level Consciousness draws inspiration from our rich human history, tracing our journey from simple pattern assessments of sunsets and sunrises to our remarkable ability to measure time with precision down to fractions of a second. This is truly amazing, and if an AI can do the same, we should be worried.

In wrapping up, this test is just my own idea. I can't say for sure if it's the perfect test for consciousness, but I believe the real answer lies somewhere in all of this. It's possible my test isn't quite right because there might be things I don't know, and you, the reader, might see that. Still, I hope that if this test is even somewhat close to the right one, we can work together to improve and use it so that we can evaluate AI that's safe for everyone.

Thank you for dedicating your time to reading this paper. The primary goal behind writing this paper was to share this idea to invite further contributions of thoughts, criticisms, or refinements. Feel free to share your insights, as this paper is intended to be a collaborative effort, and diverse perspectives will undoubtedly enhance its depth and quality. Your thoughtful engagement is greatly appreciated.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arain, M., Haque, M., Johal, L., Mathur, P., Nel, W., Rais, A., Sandhu, R., & Sharma, S. (2013). Maturation of the adolescent brain. *Neuropsychiatric Disease and Treatment*, *9*, 449–461.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Le Neindre, P., Bernard, E., Boissy, A., Boivin, X., Calandreau, L., Delon, N., Deputte, B., Desmoulin-Canselier, S., Dunier, M., Faivre, N., Giurfa, M., Guichet, J.-L., Lansade, L., Larrère, R., Mormède, P., Prunet, P., Schaal, B., Servière, J., & Terlouw, C. (2017). *Animal consciousness*, Vol. 2017:EN-1196. EFSA supporting publication.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere?. *Philosophical Transactions of the Royal Society B*, *370*(1668), 20140167.