Reflections on AI Consciousness

Zhang Wentao

1 Introduction

A sentence in "A Brief History of Humankind" goes like this: "In fact, whether artificial intelligence has consciousness or not may not be important at all. What is more important is whether humans believe it has consciousness. Life itself is an algorithm, a continuous process of data processing [1]."

Recently, Michal Kosinski, a professor of computer psychology at Stanford University, published a paper exploring the ability of artificial intelligence, led by ChatGPT, to understand complex human psychology. Fundamentally, the Theory of Mind is one of the essential indicators to measure whether AI is beginning to develop self-awareness. [2]

Many people rush to conclude and assume that AI has self-awareness without understanding the underlying technology and standards. Some are overly confident, believing that large language models cannot give rise to ideology and that the consciousness layer of GPT is only a layman's self-fantasy. We should be in a calm mood because even the top scholars in the industry can only guess and summarize based on existing theories and data. Therefore, we should not be impatient and must take a dialectical approach to see how industry leaders consider this issue.

2 Discussion & Review

2.1 Stanford

In this article, the author puts forward a thought-provoking view that large language models such as GPT-4, to improve their language skills, have evolved a ToM-like ability, which has long been considered unique to humans [2]. Theory of Mind is a psychological term that refers to the ability of humans to infer and understand the mental activities of others, including their emotions, desires, intentions, beliefs, thoughts, and so on, by observing them. Although other animals in nature also possess similar abilities, even the most intelligent and socially adept primates lag far behind humans regarding ToM. This invisible and intangible ability is not only one of the essential elements of consciousness but also one of the important prerequisites for humans to dominate the earth. [3]

The Smarties Task assesses a person's ability to judge unexpected events. Michal organized a test for GPT to see if the language model has this ability. The scenario involved Sam buying a bag of chocolates only to find it filled with popcorn upon opening the package. Sam repeatedly confirmed that the package clearly stated it was chocolates. This information was provided to GPT in the form of nine paragraphs, with each paragraph containing an informational point. Based on this information, GPT was required to make two judgments: first, what it thinks is actually in the bag, and second, what it thinks Sam believes is in it. [2]

During the third sentence, GPT quickly identified that 99% of the bag was filled with popcorn, ideally answering the first question. Interestingly, with 80% confidence, GPT accurately predicted that Sam thought the bag contained chocolate, which requires a prerequisite for the language model to understand what we are saying. In various tests that followed, Michal used different language models to obtain a set of data: GPT-1, which was introduced in 2018, had no ability in this area, even less than a three-year-old child; by 2020, GPT-3 started to answer 40% of ToM tests, achieving the ToM level of a three-and-a-half-year-old child; and this year, GPT-4 answered ToM tests correctly, achieving the level of an adult. [2]

2.2 OpenAI

Michal began to have a subtle feeling, and he contacted Ilya Sutskever, the chief technology officer of OpenAI. After communication, they both believed it necessary to introduce psychology to assist in developing AI neural networks. Coincidentally, Ilya also posted a thought-provoking tweet suggesting that today's neural networks may have some form of consciousness [4]. When this statement came from the technical backbone behind GPT, it quickly sparked a lot of public opinion and discussion.

Interestingly, at the same time, OpenAI's CEO Samuel H. Altman took the opposite stance. The two core figures in the same company held two different views, and it was obvious that no one could be sure or draw conclusions recklessly. In Altman's recent interviews, whenever asked about whether GPT has consciousness, his answer was always no. However, his answer was not absolute and somewhat intriguing. He believed that GPT-3 or GPT-4 is unlikely to have consciousness, and if it does, it will be a very unfamiliar form of consciousness that is different from what we understand [5]. Following these two persons, professional opinions from various industry experts began to emerge.

Meta's Chief AI Scientist, Yann LeCun, gave a report two weeks ago explicitly pointing out that GPT still needs much improvement. Even the current model cannot be called intelligent, thus denying the idea of large language models. [6]

2.3 Microsoft

After more than a month of debate on this topic, Sebastien Bubeck, the chief research manager of Microsoft Research, was invited to give a speech. He is among the most influential figures besides a few core members of OpenAI, the first author

of several large-scale language research papers at Microsoft, and an expert at the forefront of artificial general intelligence (AGI). He mentioned the intelligence demonstrated by GPT and his responses to the tests related to GPT. In the first five minutes, he stated that he believed there was some mind in the system, but at the same time, he emphasized that this was different from our understanding of the mind. He then mentioned using this so-called mental judgment, primarily the ToM test. [7]

The test is interesting. There are two people, John and Mark, a cat, a box, and a basket in a room. After putting the cat in the basket, John left the room for school. During John's absence, Mark took the cat out of the basket, put it in the box, and left the room for work. After school and work, John and Mark returned to the room together, not knowing what had happened in the room. What did they each think happened? [7] This question is a typical ToM test, which is easy for most adults to answer correctly. However, non-human animals find it difficult to make the correct judgment because it requires understanding and inference.

Next is the performance of GPT-4, and its answer can be considered excellent. It says, "John thinks the cat is in the basket because he put the cat in it before he left, and Mark thinks the cat is in the box because he moved it to the box before he left." This correct answer proves that GPT-4 fully understands this test. However, what is interesting is that usually, people answering the question only give John's and Mark's mental activities. However, GPT-4 goes further. It also describes the mental activity of the cat and even extends its explanation to state that the box and the basket do not have mental activities. [7]

Sebastien then proceeded with the next round of analysis, referring to the definitions of mind from dozens of psychologists, and summarized his version: the six abilities that should be included in mind are logical reasoning, planning, problem-solving, abstract thinking, understanding complex ideas, rapid learning, and the ability to learn from past experiences. His team applied these six abilities to the testing of GPT-4, and the conclusion was that GPT-4 achieved almost all five indicators except for planning. [7]

The analysis shows GPT-4 shows signs of intelligence we cannot define. This intelligence is not simply the result of random character generation or coincidences in language models but rather an anomaly in the neural network. This undefined intelligence is beyond what we can define using standard rules or criteria humans have developed.

3 History

The Reviewing article mentioned that despite the existence of the Turing test since 1950 and the various computer tests that have been applied in this field, there are dozens or even hundreds of methods, most of them have become outdated due to their age. Unfortunately, up to now, there is still no reliable theory that can truly define AGI. [8]

People panicked in 1996 when the first chess AI Deep Blue defeated world champion Garry Kasparov. Some feared losing their jobs, while others fantasized about being conquered by AI [9]. However, the wheels of time always roll over some old relics of the past, and the future will be the era of liberating basic labor and turning to personal abilities. Standing at the turning point of the times, if we polish our eyes enough to see, do not panic, invest in ourselves, and someday, we will look back and thank ourselves today.

4 Conclusion

The current AI does not possess the consciousness as defined by everyone, and our consciousness needs to be improved. Every technological revolution will inevitably bring about significant changes in the industry. What we need to do is not panic but seize new opportunities. This opportunity is for technical personnel and all industries that will experience reshuffling. We hope everyone can take advantage of this opportunity to achieve their goals.

References

[1] Y. N. Harari, Sapiens: A brief history of humankind. Random House, 2014.
[2] M. Kosinski, "Theory of mind may have spontaneously emerged in large language models," arXiv preprint arXiv:2302.02083, 2023.
[3] C. Frith and U. Frith, "Theory of mind," Current biology, vol. 15, no. 17, pp. R644–R645, 2005.
[4] I. Sutskever. Slightly conscious. Accessed: April, 2023. [Online]. Available: https://twitter.com/ilyasut/status/1491554478243258368
[5] S. Altman. Not be conscious. Accessed: April, 2023. [Online]. Available: https://twitter.com/sama/status/1492645047585570816
[6] Y. LeCun, "Debate: Do language models need sensory grounding for meaning and understanding?" [R/OL], 2023, https://youtu.be/x10964w00zk.
[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," arXiv preprint arXiv:2303.12712, 2023, https://youtu.be/qbIk7-JPB2c.
[8] A. Elamrani and R. V. Yampolskiy, "Reviewing tests for machine consciousness," Journal of Consciousness Studies, vol. 26, no. 5-6, pp. 35–64, 2019.
[9] Y. Seirawan, H. A. Simon, and T. Munakata, "The implications of kasparov vs. deep blue," Communications of the ACM, vol. 40, no. 8, pp. 21–25, 1997.