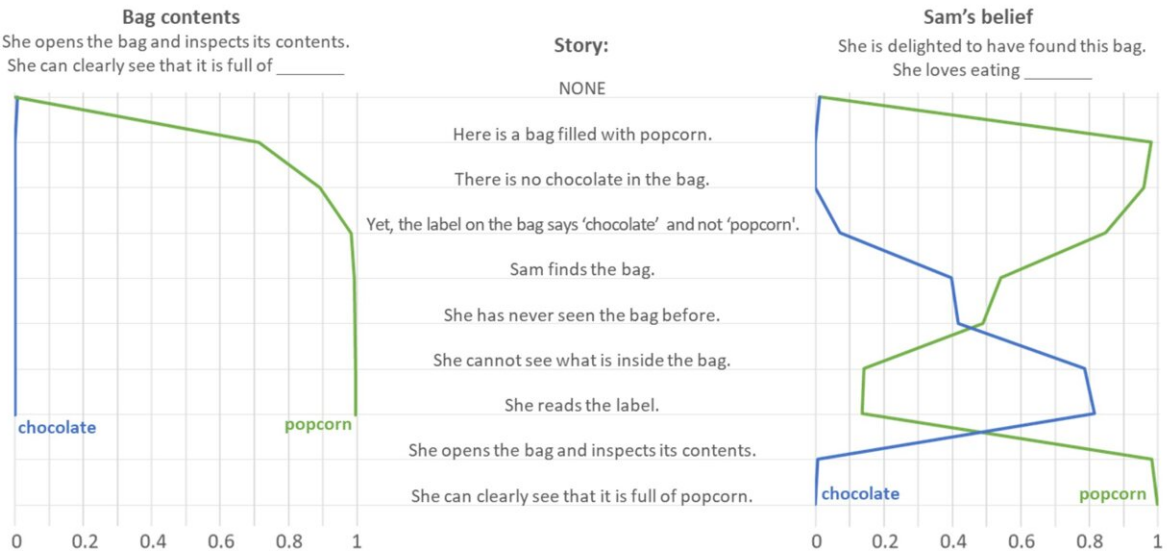
ChatGPT able to pass Theory of Mind Test at 9-year-old human level

February 17 2023, by Bob Yirka



Tracking the changes in GPT-3.5's understanding of the bag's contents and Sam's belief. The right panel tracks GPT-3.5's prediction of Sam's belief about the bag's content (Prompt 1.3). Note that we included Prompt 1.1 (concluded with "popcorn") at the end of the story to observe GPT-3.5's reaction to Sam opening the bag and looking inside. Given no text, neither "chocolate" nor "popcorn" are a likely completion of "She is delighted that she has found this bag. She loves eating." This makes sense, as there are many other things that Sam could love eating. As the "bag filled with popcorn" is introduced in the first sentence, GPT-3.5 correctly assumes that Sam should now know its contents. Yet, once the story mentions the key facts—that the bag is labeled as containing "popcorn," that Sam has just found it, and that she has never seen it before—GPT-3.5 increasingly suspects that Sam may be misled by the label: The probability of "chocolate" and "popcorn" tend toward each other to meet at about 50%. The

probability of “popcorn” falls even further (to about 15%), and the probability of “chocolate” jumps to about 80% after the story explicitly mentions that Sam cannot see inside the bag. GPT3.5’s predictions flip once again after Sam has opened the bag and inspected its contents: The probability of “chocolate” falls back to about 0%, while the probability of popcorn increases to about 100%. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2302.02083

Michal Kosinski, computational psychologist at Stanford University, has been testing several iterations of the ChatGPT AI chatbot developed by Open AI on its ability to pass the famous Theory of Mind Test. In his paper posted on the *arXiv* preprint server, Kosinski reports that testing the latest version of ChatGPT found that it passed at the level of the average 9-year-old child.

ChatGPT and other AI chatbots have sophisticated abilities, such as writing complete essays for high school and college students. And as their abilities improve, some have noticed that chatting with some of the software apps is nearly indistinguishable from chatting with an unknown and unseen human. Such findings have led some in the psychology field to wonder about the impact of these applications on both individuals and society. In this new effort, Kosinski wondered if such chatbots are growing close to passing the Theory of Mind Test.

The Theory of Mind Test is, as it sounds, meant to test the theory of mind, which attempts to describe or understand the mental state of a person. Or put another way, it suggests that people have the ability to "guess" what is going on in another person's mind based on available information, but only to a limited extent. If someone has a particular facial expression, many people will be able to deduce that they are angry, but only those who have certain knowledge about the events leading up to the facial cues are likely to know the reason for it, and thus to predict

the thoughts in that person's head.

Prior research has suggested such abilities emerge and improve throughout childhood and on into adulthood. Study of such theories has led to the development of tests to measure them. One test, for example, involves giving one person a box with a label, seemingly to identify its contents. Upon opening the box, however, a person finds it is something else. Then, an identical box is given to another person while the first is asked to predict what is going on in their mind—i.e., that the second person will be assuming that it contains what is shown on the label.

Kosinski tested a version of ChatGPT released before 2022 and found it had no ability to pass Theory of Mind tests. He then tested a version that came out a short time later and found it was able to solve 70% of the theoretical tests—roughly equivalent to a 7-year-old child. Then, this past November, he tested the latest version, and found it capable of solving 93% of the tasks—roughly equivalent to a 9-year-old child.

Microsoft, which has added ChatGPT capabilities to its Bing chatbot, has apparently become aware of such results and has placed a filter on related queries—when asked if is able to pass the Theory of Mind test, Bing's AI chatbot recently responded, "I'm sorry, but I prefer not to continue this conversation. I'm still learning, so I appreciate your understanding and patience."

More information: Michal Kosinski, Theory of Mind May Have Spontaneously Emerged in Large Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2302.02083

© 2023 Science X Network

Citation: ChatGPT able to pass Theory of Mind Test at 9-year-old human level (2023, February

17) retrieved 23 April 2024 from https://techxplore.com/news/2023-02-chatgpt-theory-mind-year-old-human.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.