The New York Times https://www.nytimes.com/2023/03/27/science/ai-machine-learning-chatbots.html

MIND

Can a Machine Know That We Know What It Knows?

Some researchers claim that chatbots have developed theory of mind. But is that just our own theory of mind gone wild?

By Oliver Whang

March 27, 2023, 11:42 a.m. ET

Mind reading is common among us humans. Not in the ways that psychics claim to do it, by gaining access to the warm streams of consciousness that fill every individual’s experience, or in the ways that mentalists claim to do it, by pulling a thought out of your head at will. Everyday mind reading is more subtle: We take in people’s faces and movements, listen to their words and then decide or intuit what might be going on in their heads.

Among psychologists, such intuitive psychology — the ability to attribute to other people mental states different from our own — is called theory of mind, and its absence or impairment has been linked to autism, schizophrenia and other developmental disorders. Theory of mind helps us communicate with and understand one another; it allows us to enjoy literature and movies, play games and make sense of our social surroundings. In many ways, the capacity is an essential part of being human.

What if a machine could read minds, too?

Recently, Michal Kosinski, a psychologist at the Stanford Graduate School of Business, made just that argument: that large language models like OpenAI's ChatGPT and GPT-4 — next-word prediction machines trained on vast amounts of text from the internet — have developed theory of mind. His studies have not been peer reviewed, but they prompted scrutiny and conversation among cognitive scientists, who have been trying to take the often asked question these days — Can ChatGPT do *this*? — and move it into the realm of more robust scientific inquiry. What capacities do these models have, and how might they change our understanding of our own minds?

"Psychologists wouldn't accept any claim about the capacities of young children just based on anecdotes about your interactions with them, which is what seems to be happening with ChatGPT," said Alison Gopnik, a psychologist at the University of California, Berkeley and one of the first researchers to look into theory of mind in the 1980s. "You have to do quite careful and rigorous tests."

Dr. Kosinski's previous research showed that neural networks trained to analyze facial features like nose shape, head angle and emotional expression could predict people's political views and sexual orientation with a startling degree of accuracy (about 72 percent in the first case and about 80 percent in the second case). His recent work on large language models uses classic theory of mind tests that measure the ability of children to attribute false beliefs to other people.

A famous example is the Sally-Anne test, in which a girl, Anne, moves a marble from a basket to a box when another girl, Sally, isn't looking. To know where Sally will look for the marble, researchers claimed, a viewer would have to exercise theory of mind, reasoning about Sally's perceptual evidence and belief formation: Sally didn't see Anne move the marble to the box, so she still believes it is where she last left it, in the basket.

Dr. Kosinski presented 10 large language models with 40 unique variations of these theory of mind tests — descriptions of situations like the Sally-Anne test, in which a person (Sally) forms a false belief. Then he asked the models questions about those situations, prodding them to see whether they would attribute false beliefs to the characters involved and accurately predict their behavior. He found that GPT-3.5, released in November 2022, did so 90 percent of the time, and GPT-4, released in March 2023, did so 95 percent of the time.

The conclusion? Machines have theory of mind.

Michal Kosinski, a psychologist at the Stanford Graduate School of Business, has argued that large language models have developed theory of mind. Many scholars disagree. Christie Hemm Klok for The New York Times

But soon after these results were released, Tomer Ullman, a psychologist at Harvard University, responded with a set of his own experiments, showing that small adjustments in the prompts could completely change the answers generated by even the most sophisticated large language models. If a container was described as transparent, the machines would fail to infer that someone could see into it. The machines had difficulty taking into account the testimony of people in these situations, and sometimes couldn't distinguish between an object being inside a container and being on top of it.

Maarten Sap, a computer scientist at Carnegie Mellon University, fed more than 1,000 theory of mind tests into large language models and found that the most advanced transformers, like ChatGPT and GPT-4, passed only about 70 percent of the time. (In other words, they were 70 percent successful at attributing false beliefs to the people described in the test situations.) The discrepancy between his data and Dr. Kosinski's could come down to differences in the testing, but Dr. Sap said that even passing 95 percent of the time would not be evidence of real theory of mind. Machines usually fail in a patterned way, unable to engage in abstract reasoning and often making "spurious correlations," he said.

Dr. Ullman noted that machine learning researchers have struggled over the past couple of decades to capture the flexibility of human knowledge in computer models. This difficulty has been a “shadow finding,” he said, hanging behind every exciting innovation. Researchers have shown that language models will often give wrong or irrelevant answers when primed with unnecessary information before a question is posed; some chatbots were so thrown off by hypothetical discussions about talking birds that they eventually claimed that birds could speak. Because their reasoning is sensitive to small changes in their inputs, scientists have called the knowledge of these machines “brittle.”

Dr. Gopnik compared the theory of mind of large language models to her own understanding of general relativity. “I have read enough to know what the words are,” she said. “But if you asked me to make a new prediction or to say what Einstein’s theory tells us about a new phenomenon, I’d be stumped because I don’t really have the theory in my head.” By contrast, she said, human theory of mind is linked with other common-sense reasoning mechanisms; it stands strong in the face of scrutiny.

In general, Dr. Kosinski’s work and the responses to it fit into the debate about whether the capacities of these machines can be compared to the capacities of humans — a debate that divides researchers who work on natural language processing. Are these machines stochastic parrots, or alien intelligences, or fraudulent tricksters? A 2022 survey of the field found that, of the 480 researchers who responded, 51 percent believed that large language models could eventually “understand natural language in some nontrivial sense,” and 49 percent believed that they could not.

Dr. Ullman doesn’t discount the possibility of machine understanding or machine theory of mind, but he is wary of attributing human capacities to nonhuman things. He noted a famous 1944 study by Fritz Heider and Marianne Simmel, in which participants were shown an animated movie of two triangles and a circle interacting. When the subjects were asked to write down what transpired in the movie, nearly all described the shapes as people.

“Lovers in the two-dimensional world, no doubt; little triangle number-two and sweet circle,” one participant wrote. “Triangle-one (hereafter known as the villain) spies the young love. Ah!”

It’s natural and often socially required to explain human behavior by talking about beliefs, desires, intentions and thoughts. This tendency is central to who we are — so central that we sometimes try to read the minds of things that don’t have minds, at least not minds like our own.