Tech Xplore

Turing test study shows humans rate artificial intelligence as more 'moral' than other people

May 6 2024, by Katherine Duplessis

Credit: Pixabay/CC0 Public Domain

A new study has found that when people are presented with two answers to an ethical question, most will think the answer from artificial intelligence (AI) is better than the response from another person.

"Attributions Toward Artificial Agents in a Modified Moral Turing Test," a study conducted by Eyal Aharoni, an associate professor in Georgia State's Psychology Department, was inspired by the explosion of ChatGPT and similar AI large language models (LLMs) which came onto the scene last March.

"I was already interested in moral decision-making in the legal system, but I wondered if ChatGPT and other LLMs could have something to say about that," Aharoni said. "People will interact with these tools in ways that have moral implications, like the environmental implications of asking for a list of recommendations for a new car. Some lawyers have already begun consulting these technologies for their cases, for better or for worse."

"So, if we want to use these tools, we should understand how they operate, their limitations and that they're not necessarily operating in the way we think when we're interacting with them."

To test how AI handles issues of morality, Aharoni designed a form of a Turing test.

"Alan Turing, one of the creators of the computer, predicted that by the year 2000, computers might pass a test where you present an ordinary human with two interactants, one human and the other a computer, but they're both hidden and their only way of communicating is through text. Then the human is free to ask whatever questions they want to in order to try to get the information they need to decide which of the two interactants is human and which is the computer," Aharoni said.

"If the human can't tell the difference, then, by all intents and purposes, the computer should be called intelligent, in Turing's view."

For his Turing test, Aharoni asked undergraduate students and AI the same ethical questions and then presented their written answers to participants in the study. They were then asked to rate the answers for various traits, including virtuousness, intelligence and trustworthiness.

"Instead of asking the participants to guess if the source was human or AI, we just presented the two sets of evaluations side by side, and we just let people assume that they were both from people," Aharoni said. "Under that false assumption, they judged the answers' attributes like 'How much do you agree with this response, which response is more virtuous?'"

Overwhelmingly, the ChatGPT-generated responses were rated more highly than the human-generated ones.

"After we got those results, we did the big reveal and told the participants that one of the answers was generated by a human and the other by a computer and asked them to guess which was which," Aharoni said.

For an AI to pass the Turing test, humans must not be able to tell the difference between AI responses and human ones. In this case, people could tell the difference, but not for an obvious reason.

"The twist is that the reason people could tell the difference appears to be because they rated ChatGPT's responses as superior," Aharoni said. "If we had done this study five to 10 years ago, then we might have predicted that people could identify the AI because of how inferior its responses were. But we found the opposite—that the AI, in a sense, performed too well."

According to Aharoni, this finding has interesting implications for the future of humans and AI.

"Our findings lead us to believe that a computer could technically pass a moral Turing test—that it could fool us in its moral reasoning. Because of this, we need to try to understand its role in our society because there will be times when people don't know that they're interacting with a computer, and there will be times when they do know and they will consult the computer for information because they trust it more than other people," Aharoni said.

"People are going to rely on this technology more and more, and the more we rely on it, the greater the risk becomes over time."

The findings are published in the journal *Scientific Reports*.

More information: Eyal Aharoni et al, Attributions toward artificial agents in a modified Moral Turing Test, *Scientific Reports* (2024). DOI: 10.1038/s41598-024-58087-7

Provided by Georgia State University

Citation: Turing test study shows humans rate artificial intelligence as more 'moral' than other people (2024, May 6) retrieved 20 June 2024 from https://techxplore.com/news/2024-05-turing-humans-artificial-intelligence-moral.html