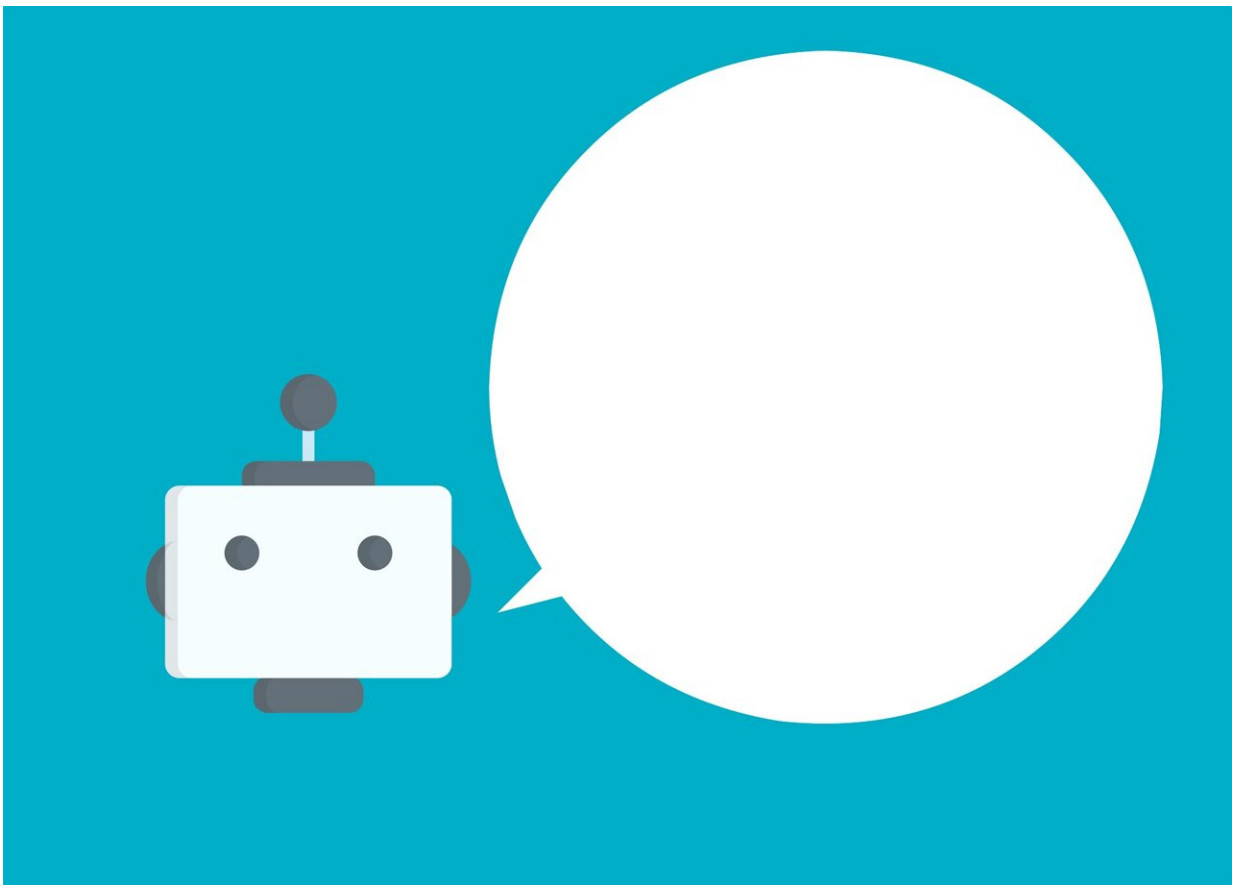
Study shows users can be primed to believe certain things about an AI chatbot's motives, influencing their interactions

October 2 2023, by Adam Zewe



Credit: Pixabay/CC0 Public Domain

Someone's prior beliefs about an artificial intelligence agent, like a

chatbot, have a significant effect on their interactions with that agent and their perception of its trustworthiness, empathy, and effectiveness, according to a new study.

Researchers from MIT and Arizona State University found that priming users—by telling them that a conversational AI agent for mental health support was either empathetic, neutral, or manipulative—influenced their perception of the chatbot and shaped how they communicated with it, even though they were speaking to the exact same chatbot.

Most users who were told the AI agent was caring believed that it was, and they also gave it higher performance ratings than those who believed it was manipulative. At the same time, less than half of the users who were told the agent had manipulative motives thought the chatbot was actually malicious, indicating that people may try to "see the good" in AI the same way they do in their fellow humans.

The study revealed a feedback loop between users' mental models, or their perception of an AI agent, and that agent's responses. The sentiment of user-AI conversations became more positive over time if the user believed the AI was empathetic, while the opposite was true for users who thought it was nefarious.

"From this study, we see that to some extent, the AI is the AI of the beholder," says Pat Pataranutaporn, a graduate student in the Fluid Interfaces group of the MIT Media Lab and co-lead author of a paper describing this study. "When we describe to users what an AI agent is, it does not just change their mental model, it also changes their behavior. And since the AI responds to the user, when the person changes their behavior, that changes the AI, as well."

Pataranutaporn is joined by co-lead author and fellow MIT graduate student Ruby Liu; Ed Finn, associate professor in the Center for Science

and Imagination at Arizona State University; and senior author Pattie Maes, professor of media technology and head of the Fluid Interfaces group at MIT.

The study, published in *Nature Machine Intelligence*, highlights the importance of studying how AI is presented to society, since the media and popular culture strongly influence our mental models. The authors also raise a cautionary flag, since the same types of priming statements in this study could be used to deceive people about an AI's motives or capabilities.

"A lot of people think of AI as only an engineering problem, but the success of AI is also a human factors problem. The way we talk about AI, even the name that we give it in the first place, can have an enormous impact on the effectiveness of these systems when you put them in front of people. We have to think more about these issues," Maes says.

AI friend or foe?

In this study, the researchers sought to determine how much of the empathy and effectiveness people see in AI is based on their subjective perception and how much is based on the technology itself. They also wanted to explore whether one could manipulate someone's subjective perception with priming.

"The AI is a black box, so we tend to associate it with something else that we can understand. We make analogies and metaphors. But what is the right metaphor we can use to think about AI? The answer is not straightforward," Pataranutaporn says.

They designed a study in which humans interacted with a conversational AI mental health companion for about 30 minutes to determine whether

they would recommend it to a friend, and then rated the agent and their experiences. The researchers recruited 310 participants and randomly split them into three groups, which were each given a priming statement about the AI.

One group was told the agent had no motives, the second group was told the AI had benevolent intentions and cared about the user's well-being, and the third group was told the agent had malicious intentions and would try to deceive users. While it was challenging to settle on only three primers, the researchers chose statements they thought fit the most common perceptions about AI, Liu says.

Half the participants in each group interacted with an AI agent based on the generative language model GPT-3, a powerful deep-learning model that can generate human-like text. The other half interacted with an implementation of the chatbot ELIZA, a less sophisticated rule-based natural language processing program developed at MIT in the 1960s.

Molding mental models

Post-survey results revealed that simple priming statements can strongly influence a user's mental model of an AI agent, and that the positive primers had a greater effect. Only 44% of those given negative primers believed them, while 88% of those in the positive group and 79% of those in the neutral group believed the AI was empathetic or neutral, respectively.

"With the negative priming statements, rather than priming them to believe something, we were priming them to form their own opinion. If you tell someone to be suspicious of something, then they might just be more suspicious in general," Liu says.

But the capabilities of the technology do play a role, since the effects

were more significant for the more sophisticated GPT-3 based conversational chatbot.

The researchers were surprised to see that users rated the effectiveness of the chatbots differently based on the priming statements. Users in the positive group awarded their chatbots higher marks for giving mental health advice, despite the fact that all agents were identical.

Interestingly, they also saw that the sentiment of conversations changed based on how users were primed. People who believed the AI was caring tended to interact with it in a more positive way, making the agent's responses more positive. The negative priming statements had the opposite effect. This impact on sentiment was amplified as the conversation progressed, Maes adds.

The results of the study suggest that because priming statements can have such a strong impact on a user's mental model, one could use them to make an AI agent seem more capable than it is—which might lead users to place too much trust in an agent and follow incorrect advice.

"Maybe we should prime people more to be careful and to understand that AI agents can hallucinate and are biased. How we talk about AI systems will ultimately have a big effect on how people respond to them," Maes says.

In the future, the researchers want to see how AI-user interactions would be affected if the agents were designed to counteract some user bias. For instance, perhaps someone with a highly positive perception of AI is given a chatbot that responds in a neutral or even a slightly negative way so the conversation stays more balanced.

They also want to use what they've learned to enhance certain AI applications, like mental health treatments, where it could be beneficial

for the user to believe an AI is empathetic. In addition, they want to conduct a longer-term study to see how a user's mental model of an AI agent changes over time.

More information: "Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness", *Nature Machine Intelligence* (2023). DOI: 10.1038/s42256-023-00720-7

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Study shows users can be primed to believe certain things about an AI chatbot's motives, influencing their interactions (2023, October 2) retrieved 2 May 2024 from https://techxplore.com/news/2023-10-users-primed-ai-chatbot-interactions.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.