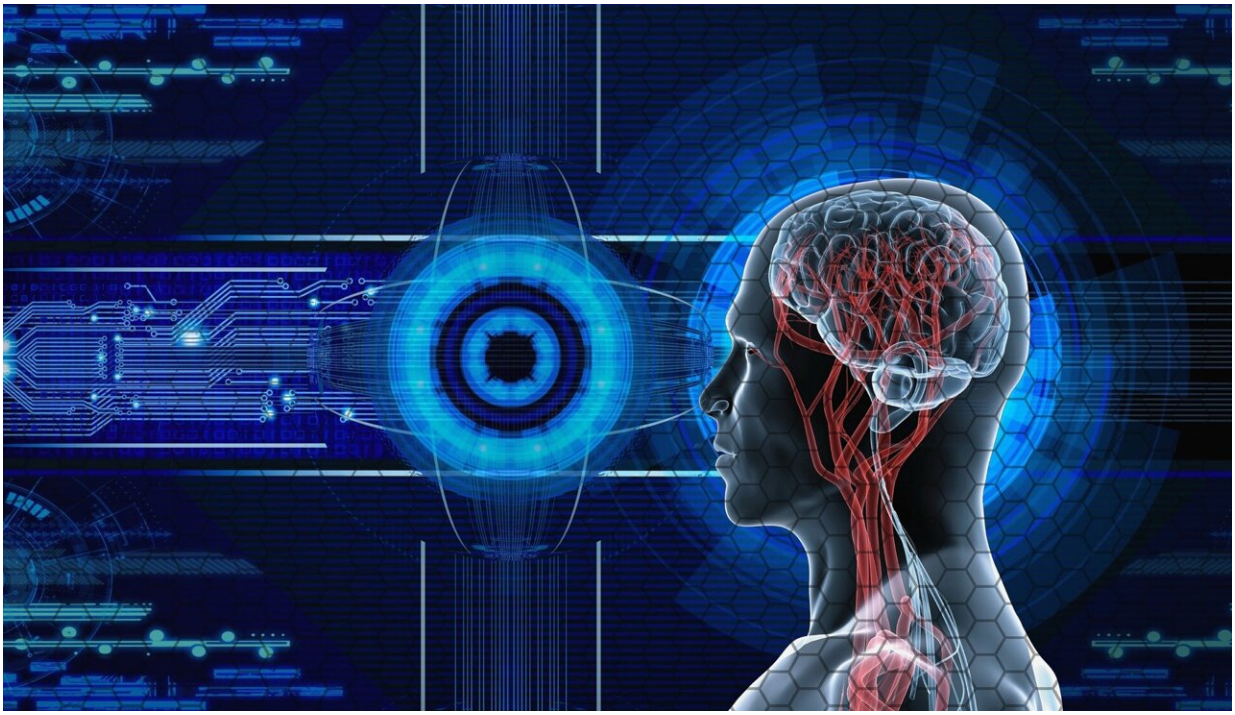
Conversations with AI can successfully reduce belief in conspiracy theories

September 12 2024



Credit: Pixabay/CC0 Public Domain

Have you ever tried to convince a conspiracy theorist that the moon landing wasn't staged? You likely didn't succeed, but ChatGPT might have better luck, according to research by MIT Sloan School of Management professor David Rand and American University professor of psychology Thomas Costello, who conducted the research during his

postdoctoral position at MIT Sloan.

In a new paper "Durably reducing conspiracy beliefs through dialogues with AI" published in *Science,* the researchers show that large language models can effectively reduce individuals' beliefs in conspiracy theories—and that these reductions last for at least two months—a finding that offers new insights into the psychological mechanisms behind the phenomenon as well as potential tools to fight the spread of conspiracies.

Going down the rabbit hole

Conspiracy theories—beliefs that certain events are the result of secret plots by influential actors—have long been a subject of fascination and concern. Their persistence in the face of counter-evidence has led to the conclusion that they fulfill deep-seated psychological needs, rendering them impervious to facts and logic. According to this conventional wisdom, once someone "falls down the rabbit hole," it's virtually impossible to pull them back out.

But for Rand, Costello, and their co-author professor Gordon Pennycook from Cornell University, who have conducted extensive research on the spread and uptake of misinformation, that conclusion didn't ring true. Instead, they suspected a simpler explanation was at play.

"We wondered if it was possible that people simply hadn't been exposed to compelling evidence disproving their theories," Rand explained. "Conspiracy theories come in many varieties—the specifics of the theory and the arguments used to support it differ from believer to believer. So if you are trying to disprove the conspiracy but haven't heard these particular arguments, you won't be prepared to rebut them."

Effectively debunking conspiracy theories, in other words, would require

two things: personalized arguments and access to vast quantities of information—both now readily available through generative AI.

Conspiracy conversations with GPT4

To test their theory, Costello, Pennycook, and Rand harnessed the power of GPT-4 Turbo, OpenAI's most advanced large language model, to engage over 2,000 conspiracy believers in personalized, evidence-based dialogues.

The study employed a unique methodology that allowed for deep engagement with participants' individual beliefs. Participants were first asked to identify and describe a conspiracy theory they believed in using their own words, along with the evidence supporting their belief.

GPT-4 Turbo then used this information to generate a personalized summary of the participant's belief and initiate a dialogue. The AI was instructed to persuade users that their beliefs were untrue, adapting its strategy based on each participant's unique arguments and evidence.

These conversations, lasting an average of 8.4 minutes, allowed the AI to directly address and refute the specific evidence supporting each individual's conspiratorial beliefs, an approach that was impossible to test at scale prior to the technology's development.

A significant—and durable—effect

The results of the intervention were striking. On average, the AI conversations reduced the average participant's belief in their chosen conspiracy theory by about 20%, and about one in four participants—all of whom believed the conspiracy beforehand—disavowed the conspiracy after the conversation. This impact proved durable, with the

effect remaining undiminished even two months post-conversation.

The AI conversation's effectiveness was not limited to specific types of conspiracy theories. It successfully challenged beliefs across a wide spectrum, including conspiracies that potentially hold strong political and social salience, like those involving COVID-19 and fraud during the 2020 U.S. presidential election.

While the intervention was less successful among participants who reported that the conspiracy was central to their worldview, it did still have an impact, with little variance across demographic groups.

Notably, the impact of the AI dialogues extended beyond mere changes in belief. Participants also demonstrated shifts in their behavioral intentions related to conspiracy theories. They reported being more likely to "unfollow" people espousing conspiracy theories online, and more willing to engage in conversations challenging those conspiratorial beliefs.

The opportunities and dangers of AI

Costello, Pennycook, and Rand are careful to point to the need for continued responsible AI deployment since the technology could potentially be used to convince users to believe in conspiracies as well as to abandon them.

Nevertheless, the potential for positive applications of AI to reduce belief in conspiracies is significant. For example, AI tools could be integrated into search engines to offer accurate information to users searching for conspiracy-related terms.

"This research indicates that evidence matters much more than we thought it did—so long as it is actually related to people's beliefs,"

Pennycook said. "This has implications far beyond just conspiracy theories: Any number of beliefs based on poor evidence could, in theory, be undermined using this approach."

Beyond the specific findings of the study, its methodology also highlights the ways in which large language models could revolutionize social science research, said Costello, who noted that the researchers used GPT-4 Turbo to not only conduct conversations but also to screen respondents and analyze data.

"Psychology research used to depend on graduate students interviewing or conducting interventions on other students, which was inherently limiting," Costello said. "Then, we moved to online survey and interview platforms that gave us scale but took away the nuance. Using artificial intelligence allows us to have both."

These findings fundamentally challenge the notion that conspiracy believers are beyond the reach of reason. Instead, they suggest that many are open to changing their views when presented with compelling and personalized counter-evidence.

"Before we had access to AI, conspiracy research was largely observation and correlational, which led to theories about conspiracies filling psychological needs," said Costello. "Our explanation is more mundane—much of the time, people just didn't have the right information."

Additionally, members of the public interested in this ongoing work can visit a website and try out the intervention for themselves.

More information: Thomas H. Costello, Durably reducing conspiracy beliefs through dialogues with AI, *Science* (2024). DOI: 10.1126/science.adq1814.

www.science.org/doi/10.1126/science.adq1814

Provided by MIT Sloan School of Management