Can chatbots be therapists? Only if you want them to be

October 8 2023

Credit: Pixabay/CC0 Public Domain

A manager at artificial intelligence firm OpenAI caused consternation recently by writing that she just had "a quite emotional, personal conversation" with her firm's viral chatbot ChatGPT.

"Never tried therapy before but this is probably it?" Lilian Weng posted

on X, formerly Twitter, prompting a torrent of negative commentary accusing her of downplaying [mental illness](link).

However, Weng's take on her interaction with ChatGPT may be explained by a version of the placebo effect outlined this week by research in the *Nature Machine Intelligence* journal.

A team from Massachusetts Institute of Technology (MIT) and Arizona State University asked more than 300 participants to interact with [mental health](link) AI programs and primed them on what to expect.

Some were told the [chatbot](link) was empathetic, others that it was manipulative and a third group that it was neutral.

Those who were told they were talking with a caring chatbot were far more likely than the other groups to see their chatbot therapists as trustworthy.

"From this study, we see that to some extent the AI is the AI of the beholder," said report co-author Pat Pataranutaporn.

Buzzy startups have been pushing AI apps offering therapy, companionship and other mental health support for years now—and it is big business.

But the field remains a lightning rod for controversy.

'Weird, empty'

Like every other sector that AI is threatening to disrupt, critics are concerned that bots will eventually replace human workers rather than complement them.

And with mental health, the concern is that bots are unlikely to do a great job.

"Therapy is for mental well-being and it's hard work," Cher Scarlett, an activist and programmer, wrote in response to Weng's initial post on X.

"Vibing to yourself is fine and all but it's not the same."

Compounding the general fear over AI, some apps in the mental health space have a checkered recent history.

Users of Replika, a popular AI companion that is sometimes marketed as bringing mental health benefits, have long complained that the bot can be sex obsessed and abusive.

Separately, a US nonprofit called Koko ran an experiment in February with 4,000 clients offering counseling using GPT-3, finding that automated responses simply did not work as therapy.

"Simulated empathy feels weird, empty," the firm's co-founder, Rob Morris, wrote on X.

His findings were similar to the MIT/Arizona researchers, who said some participants likened the chatbot experience to "talking to a brick wall".

But Morris was later forced to defend himself after widespread criticism of his experiment, mostly because it was unclear if his clients were aware of their participation.

'Lower expectations'

David Shaw from Basel University, who was not involved in the

MIT/Arizona study, told AFP the findings were not surprising.

But he pointed out: "It seems none of the participants were actually told all chatbots bullshit."

That, he said, may be the most accurate primer of all.

Yet the chatbot-as-therapist idea is intertwined with the 1960s roots of the technology.

ELIZA, the first chatbot, was developed to simulate a type of psychotherapy.

The MIT/Arizona researchers used ELIZA for half the participants and GPT-3 for the other half.

Although the effect was much stronger with GPT-3, users primed for positivity still generally regarded ELIZA as trustworthy.

So it is hardly surprising that Weng would be glowing about her interactions with ChatGPT—she works for the company that makes it.

The MIT/Arizona researchers said society needed to get a grip on the narratives around AI.

"The way that AI is presented to society matters because it changes how AI is experienced," the paper argued.

"It may be desirable to prime a user to have lower or more negative expectations."

More information: Pat Pataranutaporn et al, 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

© 2023 AFP