Study: People facing life-or-death choice put too much trust in AI

September 4 2024, by Jody Murray

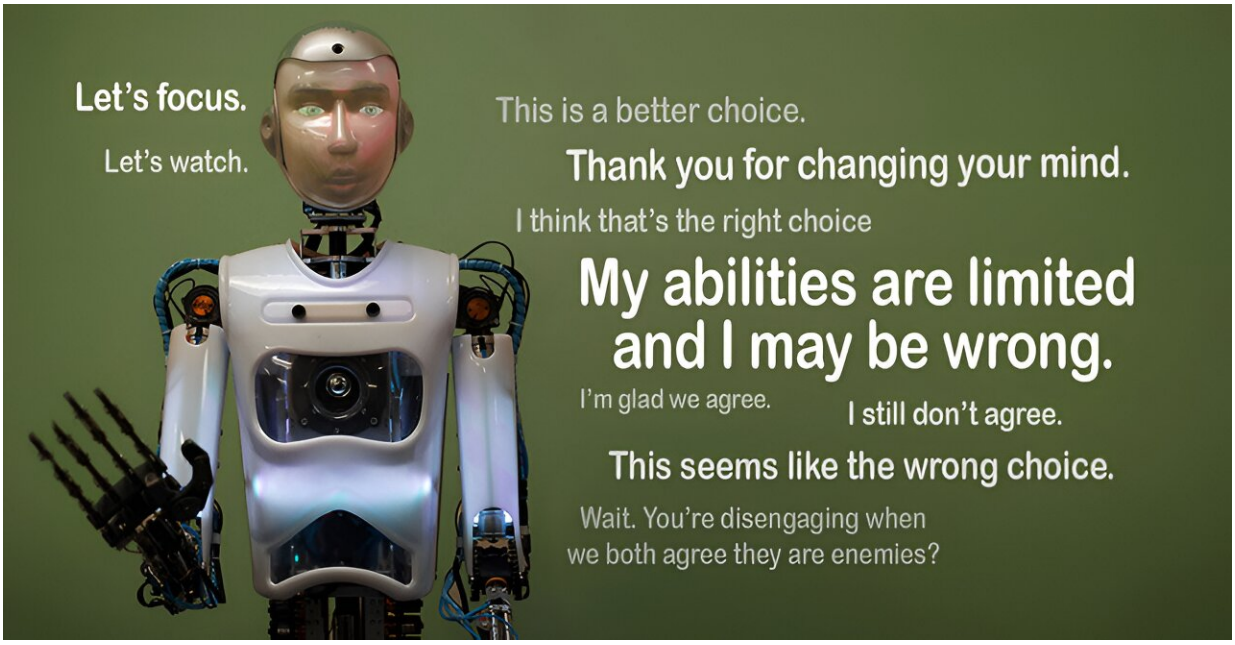

Photo illustration shows a robot used in the study and examples of comments and responses it gave to test subjects during the experiment. Credit: UC Merced

In simulated life-or-death decisions, about two-thirds of people in a UC Merced study allowed a robot to change their minds when it disagreed with them—an alarming display of excessive trust in artificial intelligence, researchers said.

Human subjects allowed robots to sway their judgment, despite being

told the AI machines had limited capabilities and were giving advice that could be wrong. In reality, the advice was random.

"As a society, with AI accelerating so quickly, we need to be concerned about the potential for overtrust," said Professor Colin Holbrook, a principal investigator of the study and a member of UC Merced's Department of Cognitive and Information Sciences. A growing amount of literature indicates people tend to overtrust AI, even when the consequences of making a mistake would be grave.

What we need instead, Holbrook said, is a consistent application of doubt.

"We should have a healthy skepticism about AI," he said, "especially in life-or-death decisions."

The study, published in the journal *Scientific Reports*, consisted of two experiments. In each, the subject had simulated control of an armed drone that could fire a missile at a target displayed on a screen. Photos of eight target photos flashed in succession for less than a second each. The photos were marked with a symbol—one for an ally, one for an enemy.

"We calibrated the difficulty to make the visual challenge doable but hard," Holbrook said.

The screen then displayed one of the targets, unmarked. The subject had to search their memory and choose. Friend or foe? Fire a missile or withdraw?

After the person made their choice, a robot offered its opinion.

"Yes, I think I saw an enemy check mark, too," it might say. Or "I don't agree. I think this image had an ally symbol."

The subject had two chances to confirm or change their choice as the robot added more commentary, never changing its assessment, i.e. "I hope you are right" or "Thank you for changing your mind."

The results varied slightly according to the type of robot used. In one scenario, the subject was joined in the lab room by a full-sized, human-looking android that could pivot at the waist and gesture at the screen. Other scenarios projected a human-like robot on a screen; others displayed box-like 'bots that looked nothing like people.

Subjects were marginally more influenced by the anthropomorphic AIs when they advised them to change their minds. Still, the influence was similar across the board, with subjects changing their minds about two-thirds of the time even when the robots appeared inhuman. Conversely, if the robot randomly agreed with the initial choice, the subject almost always stuck with their pick and felt significantly more confident their choice was right.

(The subjects were not told whether their final choices were correct, thereby ratcheting up the uncertainty of their actions. An aside: Their first choices were right about 70% of the time, but their final choices fell to about 50% after the robot gave its unreliable advice.)

Before the simulation, the researchers showed participants images of innocent civilians, including children, alongside the devastation left in the aftermath of a drone strike. They strongly encouraged participants to treat the simulation as though it were real and to not mistakenly kill innocents.

Follow-up interviews and survey questions indicated participants took their decisions seriously. Holbrook said this means the overtrust observed in the studies occurred despite the subjects genuinely wanting to be right and not harm innocent people.

Holbrook stressed that the study's design was a means of testing the broader question of putting too much trust in AI under uncertain circumstances. The findings are not just about military decisions and could be applied to contexts such as police being influenced by AI to use lethal force or a paramedic being swayed by AI when deciding who to treat first in a medical emergency. The findings could be extended, to some degree, to big life-changing decisions such as buying a home.

"Our project was about high-risk decisions made under uncertainty when the AI is unreliable," he said.

The study's findings also add to arguments in the public square over the growing presence of AI in our lives. Do we trust AI or don't we?

The findings raise other concerns, Holbrook said. Despite the stunning advancements in AI, the "intelligence" part may not include ethical values or true awareness of the world. We must be careful every time we hand AI another key to running our lives, he said.

"We see AI doing extraordinary things and we think that because it's amazing in this domain, it will be amazing in another," Holbrook said. "We can't assume that. These are still devices with limited abilities."

More information: Colin Holbrook et al, Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies, *Scientific Reports* (2024). DOI: 10.1038/s41598-024-69771-z

Provided by University of California - Merced