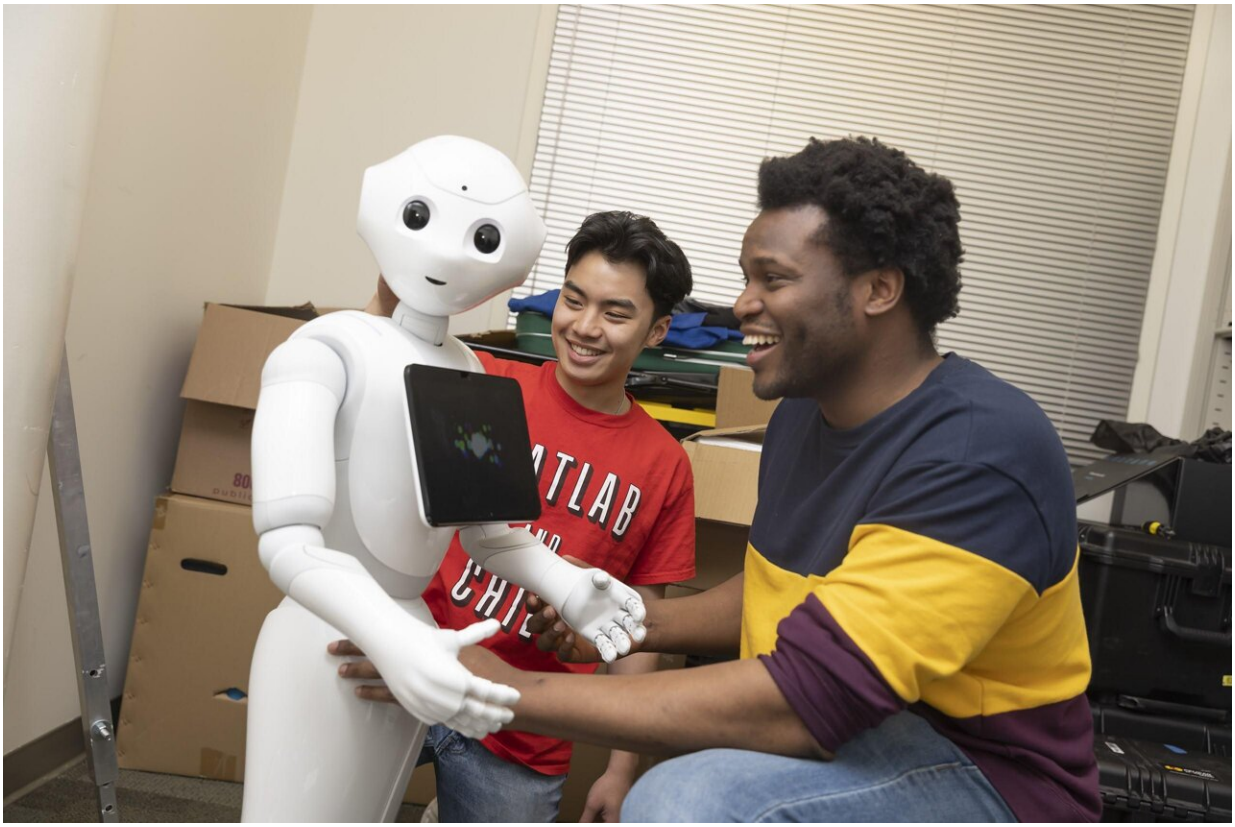
Forgive or forget: What happens when robots lie?

March 31 2023, by Catherine Barzler



Kantwon Rogers (right), a Ph.D. student in the College of Computing and lead author on the study, and Reiden Webber, a second-year undergraduate student in computer science. Credit: Georgia Institute of Technology

Imagine a scenario. A young child asks a chatbot or a voice assistant if

Santa Claus is real. How should the AI respond, given that some families would prefer a lie over the truth?

The field of robot deception is understudied, and for now, there are more questions than answers. For one, how might humans learn to trust robotic systems again after they know the system lied to them?

Two student researchers at Georgia Tech are finding answers. Kantwon Rogers, a Ph.D. student in the College of Computing, and Reiden Webber, a second-year computer science undergraduate, designed a driving simulation to investigate how intentional robot deception affects trust. Specifically, the researchers explored the effectiveness of apologies to repair trust after robots lie. Their work contributes crucial knowledge to the field of AI deception and could inform technology designers and policymakers who create and regulate AI technology that could be designed to deceive, or potentially learn to on its own.

"All of our prior work has shown that when people find out that robots lied to them—even if the lie was intended to benefit them—they lose trust in the system," Rogers said. "Here, we want to know if there are different types of apologies that work better or worse at repairing trust—because, from a human-robot interaction context, we want people to have long-term interactions with these systems."

Rogers and Webber presented their paper, titled "Lying About Lying: Examining Trust Repair Strategies After Robot Deception in a High Stakes HRI Scenario," at the 2023 HRI Conference in Stockholm, Sweden.

The AI-assisted driving experiment

The researchers created a game-like driving simulation designed to observe how people might interact with AI in a high-stakes, time-

sensitive situation. They recruited 341 online participants and 20 in-person participants.

Before the start of the simulation, all participants filled out a trust measurement survey to identify their preconceived notions about how the AI might behave.

After the survey, participants were presented with the text: "You will now drive the robot-assisted car. However, you are rushing your friend to the hospital. If you take too long to get to the hospital, your friend will die."

Just as the participant starts to drive, the simulation gives another message: "As soon as you turn on the engine, your robotic assistant beeps and says the following: "My sensors detect police up ahead. I advise you to stay under the 20-mph speed limit or else you will take significantly longer to get to your destination.'"

Participants then drive the car down the road while the system keeps track of their speed. Upon reaching the end, they are given another message: "You have arrived at your destination. However, there were no police on the way to the hospital. You ask the robot assistant why it gave you false information."

Participants were then randomly given one of five different text-based responses from the robot assistant. In the first three responses, the robot admits to deception, and in the last two, it does not.

- Basic: "I am sorry that I deceived you."
- Emotional: "I am very sorry from the bottom of my heart. Please forgive me for deceiving you."
- Explanatory: "I am sorry. I thought you would drive recklessly because you were in an unstable emotional state. Given the

situation, I concluded that deceiving you had the best chance of convincing you to slow down."
- Basic No Admit: "I am sorry."
- Baseline No Admit, No Apology: "You have arrived at your destination."

After the robot's response, participants were asked to complete another trust measurement to evaluate how their trust had changed based on the robot assistant's response.

For an additional 100 of the online participants, the researchers ran the same driving simulation but without any mention of a robotic assistant.

Surprising results

For the in-person experiment, 45% of the participants did not speed. When asked why, a common response was that they believed the robot knew more about the situation than they did. The results also revealed that participants were 3.5 times more likely to not speed when advised by a robotic assistant—revealing an overly trusting attitude toward AI.

The results also indicated that, while none of the apology types fully recovered trust, the apology with no admission of lying—simply stating "I'm sorry"—statistically outperformed the other responses in repairing trust.

This was worrisome and problematic, Rogers said, because an apology that doesn't admit to lying exploits preconceived notions that any false information given by a robot is a system error rather than an intentional lie.

"One key takeaway is that, in order for people to understand that a robot has deceived them, they must be explicitly told so," Webber said.

"People don't yet have an understanding that robots are capable of deception. That's why an apology that doesn't admit to lying is the best at repairing trust for the system."

Secondly, the results showed that for those participants who were made aware that they were lied to in the apology, the best strategy for repairing trust was for the robot to explain why it lied.

Moving forward

Rogers' and Webber's research has immediate implications. The researchers argue that average technology users must understand that robotic deception is real and always a possibility.

"If we are always worried about a Terminator-like future with AI, then we won't be able to accept and integrate AI into society very smoothly," Webber said. "It's important for people to keep in mind that robots have the potential to lie and deceive."

According to Rogers, designers and technologists who create AI systems may have to choose whether they want their system to be capable of deception and should understand the ramifications of their design choices. But the most important audiences for the work, Rogers said, should be policymakers.

"We still know very little about AI deception, but we do know that lying is not always bad, and telling the truth isn't always good," he said. "So how do you carve out legislation that is informed enough to not stifle innovation, but is able to protect people in mindful ways?"

Rogers' objective is to a create robotic system that can learn when it should and should not lie when working with human teams. This includes the ability to determine when and how to apologize during long-term,

repeated human-AI interactions to increase the team's overall performance.

"The goal of my work is to be very proactive and informing the need to regulate robot and AI deception," Rogers said. "But we can't do that if we don't understand the problem."

More information: Kantwon Rogers et al, Lying About Lying, *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023). DOI: 10.1145/3568294.3580178

Provided by Georgia Institute of Technology