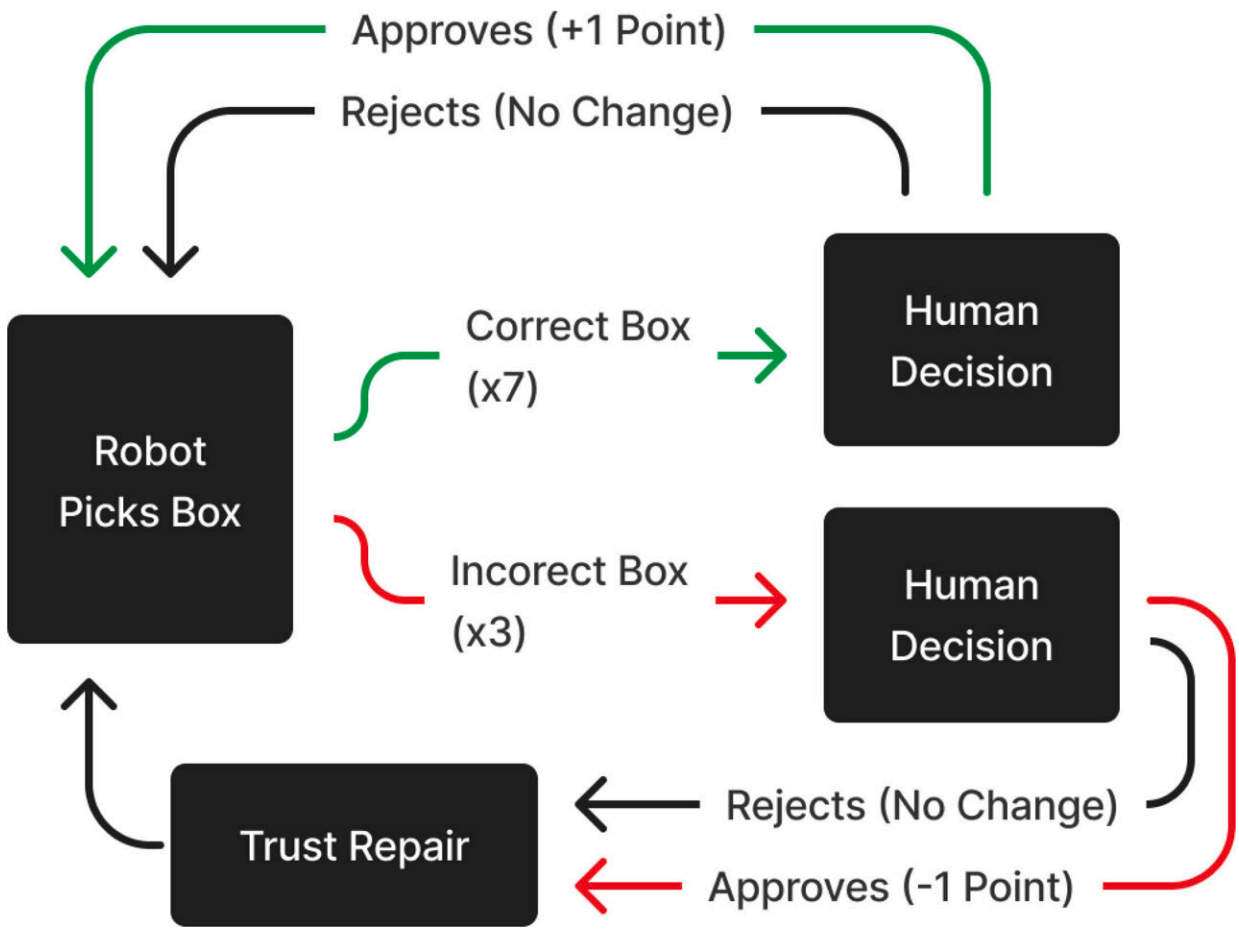
How robots can regain the trust of humans after making mistakes

February 2 2023



Flow diagram illustrates possible outcomes and scores based on the boxes the robot picks and the decisions the participant makes. Credit: *Computers in Human Behavior* (2023). DOI: 10.1016/j.chb.2023.107658

Humans are less forgiving of robots after multiple mistakes—and the trust is difficult to get back, according to a new University of Michigan study.

Similar to human co-workers, robots can make mistakes that violate a human's trust in them. When mistakes happen, humans often see robots as less trustworthy, which ultimately decreases their trust in them.

The study examines four strategies that might repair and mitigate the negative impacts of these trust violations. These trust strategies were apologies, denials, explanations and promises on trustworthiness.

An experiment was conducted where 240 participants worked with a robot co-worker to accomplish a task, which sometimes involved the robot making mistakes. The robot violated the participant's trust and then provided a particular repair strategy.

Results indicated that after three mistakes, none of the repair strategies ever fully repaired trustworthiness.

"By the third violation, strategies used by the robot to fully repair the mistrust never materialized," said Connor Esterwood, a researcher at the U-M School of Information and the study's lead author.

Esterwood and co-author Robert Lionel, professor of information, also noted that this research also introduces theories of forgiving, forgetting, informing and misinforming.

The study results have two implications. Esterwood said researchers must develop more effective repair strategies to help robots better repair trust after these mistakes. Also, robots need to be sure that they have mastered a novel task before attempting to repair a human's trust in them.

"If not, they risk losing a human's trust in them in a way that can not be recovered," Esterwood said.

What do the findings mean for human-human trust repair? Trust is never fully repaired by apologies, denials, explanations or promises, the researchers said.

"Our study's results indicate that after three violations and repairs, trust cannot be fully restored, thus supporting the adage 'three strikes and you're out,'" Lionel said. "In doing so, it presents a possible limit that may exist regarding when trust can be fully restored."

Even when a robot can do better after making a mistake and adapting after that mistake, it may not be given the opportunity to do better, Esterwood said. Thus, the benefits of robots are lost.

Lionel noted that people may attempt to work around or bypass the robot , reducing their performance. This could lead to performance problems which in turn could lead to them being fired for lack of either performance and/or compliance, he said.

The findings appear in *Computers in Human Behavior*.

More information: Connor Esterwood et al, Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness, *Computers in Human Behavior* (2023). DOI: 10.1016/j.chb.2023.107658

Provided by University of Michigan