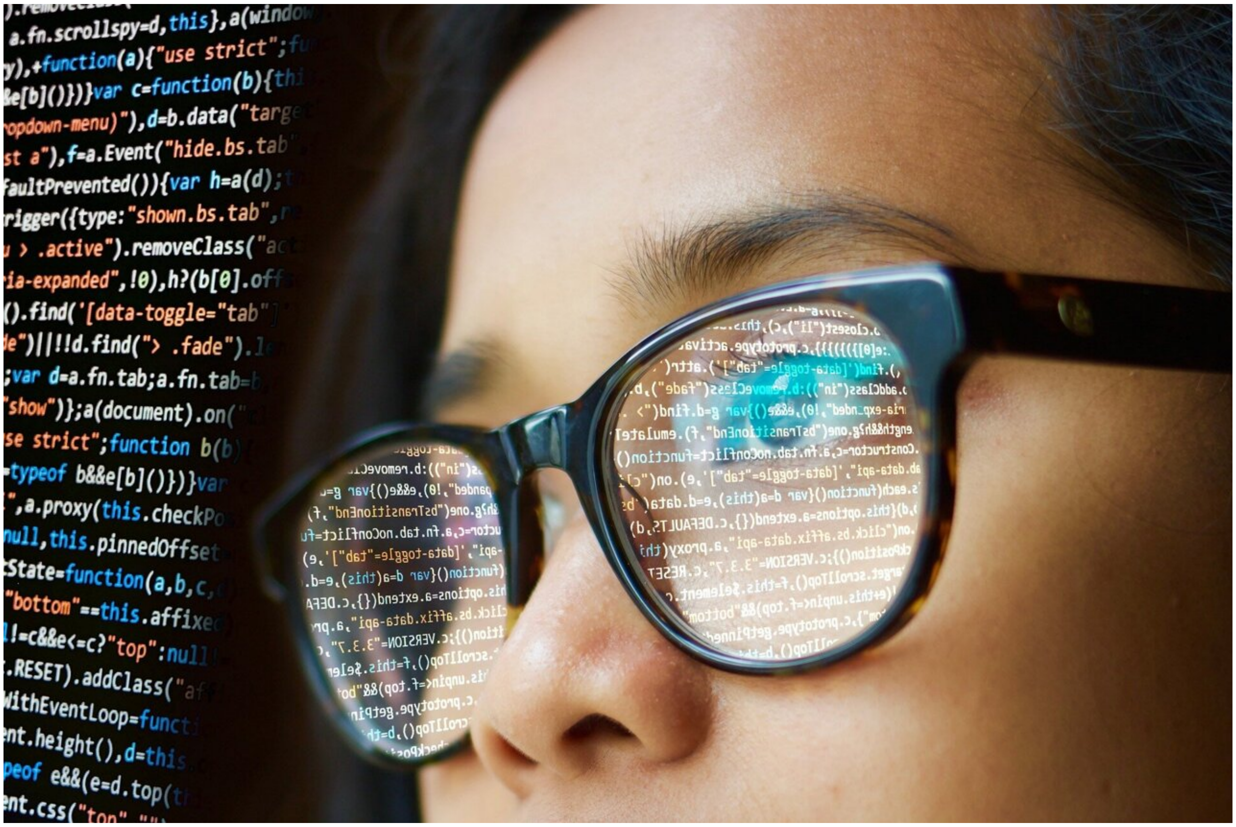
Humans change their own behavior when training AI

August 7 2024, by Chris Woolston



Credit: Pixabay/CC0 Public Domain

A new cross-disciplinary study by Washington University in St. Louis researchers has uncovered an unexpected psychological phenomenon at the intersection of human behavior and artificial intelligence: When told they were training AI to play a bargaining game, participants actively adjusted their own behavior to appear more fair and just, an impulse with potentially important implications for real-world AI developers.

"The participants seemed to have a motivation to train AI for fairness, which is encouraging, but other people might have different agendas," said Lauren Treiman, a Ph.D. student in the Division of Computational and Data Sciences and lead author of the study. "Developers should know that people will intentionally change their behavior when they know it will be used to train AI."

The study is published in *Proceedings of the National Academy of Sciences*. The co-authors are Wouter Kool, assistant professor of psychological and brain sciences in Arts & Sciences, and Chien-Ju Ho, assistant professor of computer science and engineering in the McKelvey School of Engineering. Kool and Ho are Treiman's graduate advisors.

The study included five experiments, each with roughly 200–300 participants. Subjects were asked to play the "Ultimatum Game," a challenge that requires them to negotiate small cash payouts (just $1 to $6) with other human players or a computer. In some cases, they were told their decisions would be used to teach an AI bot how to play the game.

The players who thought they were training AI were consistently more likely to seek a fair share of the payout, even if such fairness cost them a few bucks. Interestingly, that behavior change persisted even after they were told their decisions were no longer being used to train AI, suggesting the experience of shaping technology had a lasting impact on decision-making.

"As cognitive scientists, we're interested in habit formation," Kool said. "This is a cool example because the behavior continued even when it was not called for anymore."

Still, the impulse behind the behavior isn't entirely clear. Researchers didn't ask about specific motivations and strategies, and Kool explained that participants may not have felt a strong obligation to make AI more ethical. It's possible, he said, that the experiment simply brought out their natural tendencies to reject offers that seemed unfair.

"They may not really be thinking about the future consequences," he said. "They could just be taking the easy way out."

"The study underscores the important human element in the training of AI," said Ho, a computer scientist who studies the relationships between human behaviors and machine learning algorithms. "A lot of AI training is based on human decisions," he said. "If human biases during AI training aren't taken into account, the resulting AI will also be biased. In the last few years, we've seen a lot of issues arising from this sort of mismatch between AI training and deployment."

Some facial recognition software, for example, is less accurate at identifying people of color, Ho said. "That's partly because the data used to train AI is biased and unrepresentative," he said.

Treiman is now conducting follow-up experiments to get a better sense of the motivations and strategies of people training AI. "It's very important to consider the psychological aspects of computer science," she said.

More information: Lauren S. Treiman et al, The consequences of AI training on human decision-making, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2408731121

Provided by Washington University in St. Louis

Citation: Humans change their own behavior when training AI (2024, August 7) retrieved 7 August 2024 from https://techxplore.com/news/2024-08-humans-behavior-ai.html