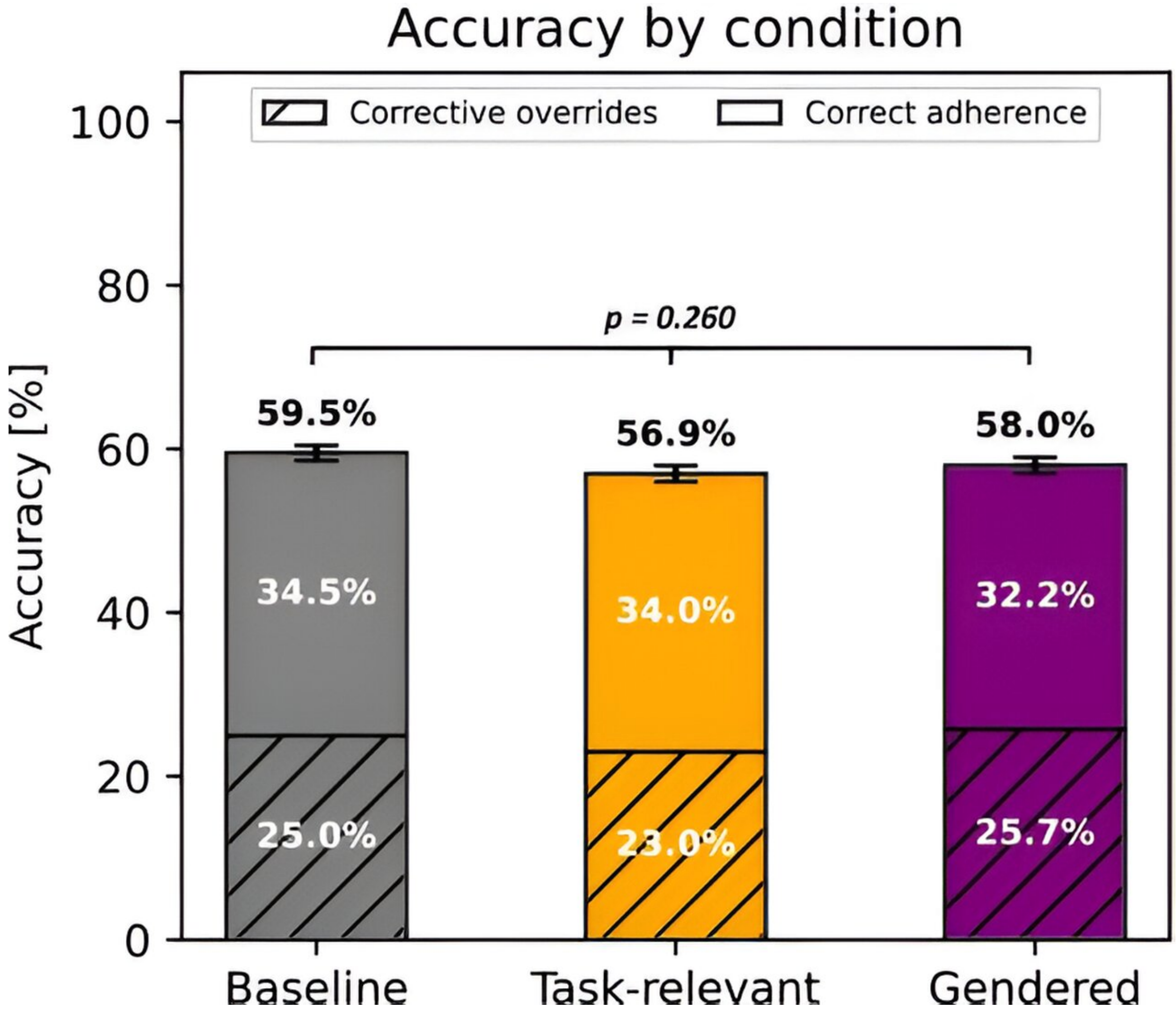
When AI aids decisions, when should humans override?

August 8 2024



Accuracy is not higher when explanations are provided, compared to the baseline. Credit: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). DOI: 10.1145/3613904.3642621

The $184 billion market for artificial intelligence shows no signs of slowing. A big slice of that market is organizations, from businesses to government agencies, that rely on AI to help make decisions. A 2023 study by IBM found 43% of CEOs are using AI to make strategic decisions.

But relying on AI can be problematic, given its well-documented history of bias, including stereotyping by race and gender. That can lead to flawed recommendations and unfair treatment when AI considers demographics to discourage granting a bank loan or a job interview to certain people.

One way to mitigate these problems, experts say, is to make AI systems explain themselves. By reviewing an AI's "explanation" of how they make decisions, human hiring managers or loan officers, for example, can decide whether to override AI recommendations.

But new research from Texas McCombs finds that the explanations themselves can be problematic. They may fuel a perception of fairness without being grounded in accuracy or equity.

"What we find is that the process doesn't lead humans to actually make better quality decisions or fairer decisions," says Maria De-Arteaga, assistant professor of information, risk, and operations management.

In the study, De-Arteaga and her co-authors—UT postdoctoral research fellow Jakob Schoeffer and Niklas Kühl of the University of Bayreuth, Germany—had an AI system read 134,436 online biographies and predict whether each person was a teacher or professor.

Then, human participants were allowed to read the bios and choose whether to override AI recommendations. There were two types of explanations, and each participant saw one of the two:

- Explanations highlighting task-relevant keywords such as "research" or "schools."
- Explanations highlighting keywords related to gender, such as "he" or "she."

The research found that participants were 4.5 percentage points more likely to override AI recommendation when the explanations highlighted gender rather than task-relevance.

A major reason: suspected gender bias. Participants were more likely to think recommendations were unfair when they focused on gender.

Illusion of accuracy and fairness

But the participants were not always correct. When it came to identifying professors or teachers, gender-based overrides were no more accurate than task-based overrides. In fact, neither type of explanation improved human accuracy, compared with participants who were not given explanations.

Why didn't the explanations lead to better decisions? De-Arteaga focuses on the participants who saw task-relevant words and assumed they were free of gender bias.

But research has found the opposite: An algorithm can develop gender bias by learning correlations between seemingly task-relevant words and gender. The explanations don't reveal that kind of bias. Humans wrongly assume the AI is gender neutral, and they decline to override it.

"There's this hope that explanations are going to help humans discern whether a recommendation is wrong or biased," De-Arteaga says. "But there's a disconnect between what the explanations are doing and what we wish they did."

Although AI explanations that try to approximate the importance of certain factors can be inherently flawed, De-Arteaga and her co-authors suggest several ways to make their design and deployment more useful to decision-makers.

- Set more concrete and realistic objectives for explanations, based on the decisions to be made, and evaluate whether they accomplish the desired goal.
- Provide more relevant cues in the explanations, such as those related to fairness in the AI system.
- Widen the scope of explanations by giving more insight into how the algorithm works. For example, knowing what data is and is not available to the system may better empower humans to use algorithms well.
- Study the psychological mechanisms at play when humans do or do not decide to override an AI decision. Recommendations should be designed to reflect how humans actually interact with AI, rather than how researchers wish they interacted.

The goal, she says, is to develop tools that help humans successfully complement AI systems—not just offer explanations that build a false sense of trust.

"That's one of the problems with explanations," she says. "They can lead you to trust the system more, even if the system is not deserving of your trust."

"Explanations, Fairness, and Appropriate Reliance in Human-AI Decision Making" is published in *Proceedings of the CHI Conference on*

Human Factors in Computing Systems.

More information: Jakob Schoeffer et al, Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making, *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). DOI: 10.1145/3613904.3642621

Provided by University of Texas at Austin

Citation: When AI aids decisions, when should humans override? (2024, August 8) retrieved 8 August 2024 from https://techxplore.com/news/2024-08-ai-aids-decisions-humans-override.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.