Lie-detection AI could provoke people into making careless accusations, researchers warn

June 27 2024

Credit: Unsplash/CC0 Public Domain

Although people lie a lot, they typically refrain from accusing others of lying because of social norms around making false accusations and being

polite. But artificial intelligence (AI) could soon shake up the rules.

In a study published June 27 in the journal *iScience*, researchers demonstrate that people are much more likely to accuse others of lying when an AI makes an accusation. The finding provides insights into the social implications of using AI systems for lie detection, which could inform policymakers when implementing similar technologies.

"Our society has strong, well-established norms about accusations of lying," says senior author Nils Köbis, a behavioral scientist at the University Duisburg-Essen in Germany.

"It would take a lot of courage and evidence for one to openly accuse others of lying. But our study shows that AI could become an excuse for people to conveniently hide behind, so that they can avoid being held responsible for the consequences of accusations."

Human society has long operated based on the truth-default theory, which explains that people generally assume what they hear is true. Because of this tendency to trust others, humans are terrible at detecting lies. Previous research has shown that people perform no better than chance when trying to detect lies.

Köbis and his team wanted to know whether the presence of AI would change the established social norms and behaviors about making accusations.

To investigate, the team asked 986 people to write one true and one false description of what they plan to do next weekend. The team then trained an algorithm with the data to develop an AI model that was able to correctly identify true and false statements 66% of the time, an accuracy significantly higher than what an average person can achieve.

Next, the team recruited more than 2,000 people to be the judges who would read a statement and decide if it is true or false. The researchers divided the participants into four groups—"baseline," "forced," "blocked," and "choice."

In the baseline group, participants answered true or false without help from the AI. In the forced group, the participants always received an AI prediction before making their own judgment. In the blocked and choice groups, participants had the option of receiving an AI-generated prediction. People who requested the prediction from the blocked group would not receive it, while people in the choice group would.

The research team found participants in the baseline group had an accuracy of 46% when identifying the statements of being true or false. Only 19% of the people in the group accused the statements they read being false, even though they knew that 50% of the statements were false. This confirms that people tend to refrain from accusing others of lying.

In the forced group where participants were given an AI prediction regardless of whether they wanted it, over a third of participants accused the statements of being false. The rate is significantly higher than both the baseline and blocked groups that received no AI predictions.

When the AI predicted a statement was true, only 13% of participants said the statement was false. However, when the AI predicted a statement as false, more than 40% of participants accused the statement of being false.

Moreover, among the participants who requested and received an AI prediction, an overwhelming 84% of them adopted the prediction and made accusations when the AI said the statement was false.

"It shows that once people have such an algorithm on hand, they would rely on it and maybe change their behaviors. If the algorithm calls something a lie, people are willing to jump on that. This is quite alarming, and it shows we should be really careful with this technology," Köbis says.

Interestingly, people seemed to be reluctant to use AI as a lie-detection tool. In the blocked and choice groups, only a third of participants requested the AI prediction.

The result was surprising to the team, because the researchers had told the participants in advance that the algorithm could detect lies better than humans. "It might be because of this very robust effect we've seen in various studies that people are overconfident in their lie detection abilities, even though humans are really bad at it," Köbis says.

AI is known for making frequent mistakes and reinforcing biases. Given the findings, Köbis suggests that policymakers should reconsider using the technology on important and sensitive matters like granting asylum at the borders.

"There's such a big hype around AI, and many people believe these algorithms are really, really potent and even objective. I'm really worried that this would make people over-rely on it, even when it doesn't work that well," Köbis says.

More information: Lie Detection Algorithms Disrupt the Social Dynamics of Accusation Behavior, *iScience* (2024). DOI: 10.1016/j.isci.2024.110201. www.cell.com/iscience/fulltext … 2589-0042(24)01426-3

Provided by Cell Press

Citation: Lie-detection AI could provoke people into making careless accusations, researchers warn (2024, June 27) retrieved 30 June 2024 from https://phys.org/news/2024-06-ai-provoke-people-careless-accusations.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.