Research shows humans can inherit AI biases

October 3 2023

Credit: CC0 Public Domain

New research by psychologists Lucía Vicente and Helena Matute from Deusto University in Bilbao, Spain, provides evidence that people can inherit artificial intelligence biases (systematic errors in AI outputs) in their decisions.

The astonishing results achieved by AI systems that can, for example, hold a conversation as a human does have given this technology an image of high reliability.

More and more professional fields are implementing AI-based tools to support the decision-making of specialists to minimize errors in their decisions. However, this technology is not without risks due to biases in AI results. We must consider that the data used to train AI models reflects past human decisions. If this data hides patterns of systematic errors, the AI algorithm will learn and reproduce these errors. Indeed, extensive evidence indicates that AI systems do inherit and amplify human biases.

The most relevant finding of Vicente and Matute's research is that the opposite effect may also occur: that humans inherit AI biases. That is, not only would AI inherit its biases from human data, but people could also inherit those biases from AI, with the risk of getting trapped in a dangerous loop. *Scientific Reports* has published the results of Vicente and Matute's research.

In the series of three experiments conducted by the researchers, volunteers performed a medical diagnosis task. A group of the participants were assisted by a biased AI system (it exhibited a systematic error) during this task, while the control group were unassisted. The AI, the medical diagnosis task, and the disease were fictitious. The whole setting was a simulation to avoid interference with real situations.

The participants assisted by the biased AI system made the same type of errors as the AI, while the control group did not make these mistakes. Thus, AI recommendations influenced participant's decisions.

Yet the most significant finding of the research was that after interaction

with the AI system, those volunteers continued to mimic its systematic error when they switched to performing the diagnosis task unaided. In other words, participants who were first assisted by the biased AI replicated its bias in a context without this support, thus showing an inherited bias. This effect was not observed for the participants in the control group, who performed the task unaided from the beginning.

These results show that biased information by an artificial intelligence model can have a perdurable negative impact on human decisions. The finding of an inheritance of AI bias effect points to the need for further psychological and multidisciplinary research on AI-human interaction.

Furthermore, evidence-based regulation is also needed to guarantee fair and ethical AI, considering not only the AI technical features but also the psychological aspects of the AI and human collaboration.

More information: Lucía Vicente et al, Humans inherit artificial intelligence biases, *Scientific Reports* (2023). DOI: 10.1038/s41598-023-42384-8

Provided by University of Deusto