Bias in AI amplifies our own biases, finds study

December 18 2024

Credit: Pixabay/CC0 Public Domain

Artificial intelligence (AI) systems tend to take on human biases and amplify them, causing people who use that AI to become more biased themselves, finds a new study by UCL researchers.

Human and AI biases can consequently create a feedback loop, with small initial biases increasing the risk of human error, according to the

findings published in *Nature Human Behaviour*.

The researchers demonstrated that AI bias can have real-world consequences, as they found that people interacting with biased AIs became more likely to underestimate women's performance and overestimate white men's likelihood of holding high-status jobs.

Co-lead author Professor Tali Sharot (UCL Psychology & Language Sciences, Max Planck UCL Center for Computational Psychiatry and Aging Research, and Massachusetts Institute of Technology) said, "People are inherently biased, so when we train AI systems on sets of data that have been produced by people, the AI algorithms learn the human biases that are embedded in the data. AI then tends to exploit and amplify these biases to improve its prediction accuracy.

"Here, we've found that people interacting with biased AI systems can then become even more biased themselves, creating a potential snowball effect wherein minute biases in original datasets become amplified by the AI, which increases the biases of the person using the AI."

The researchers conducted a series of experiments with over 1,200 study participants who were completing tasks and interacting with AI systems.

As a precursor to one of the experiments, the researchers trained an AI algorithm on a dataset of participant responses. People were asked to judge whether a group of faces in a photo looked happy or sad, and they demonstrated a slight tendency to judge faces as sad more often than happy. The AI learned this bias and amplified it into a greater bias towards judging faces as sad.

Another group of participants then completed the same task, but were also told what judgment the AI had made for each photo.

After interacting with this AI system for a while, this group of people internalized the AI's bias and were even more likely to say faces looked sad than before interacting with the AI. This demonstrates that the AI learned a bias from a human-derived dataset, and then amplified the inherent biases of another group of people.

The researchers found similar results in experiments using very different tasks, including assessing the direction a set of dots was moving across a screen, or, notably, assessing another person's performance on a task, wherein people were particularly likely to overestimate men's performance after interacting with a biased AI system (which was created with an inherent gender bias to imitate the biases of many existing AIs). The participants were generally unaware of the extent of AI influence.

When people were falsely told they were interacting with another person, but in truth were interacting with an AI, they internalized the biases to a lesser extent, which the researchers say could be because people expect AI to be more accurate than a human on some tasks.

The researchers also conducted experiments with a widely-used generative AI system, Stable Diffusion.

In one experiment, the researchers prompted the AI to generate photos of financial managers, which yielded biased results, as white men were overrepresented beyond their actual share.

They then asked study participants to view a series of headshots and select which person is most likely to be a financial manager before and after being presented with the images generated by the AI. The researchers found participants were even more inclined to indicate a white man was most likely to be a financial manager after viewing the images generated by Stable Diffusion than before.

Co-lead author Dr. Moshe Glickman (UCL Psychology & Language Sciences and Max Planck UCL Center for Computational Psychiatry and Aging Research) said, "Not only do biased people contribute to biased AIs, but biased AI systems can alter people's own beliefs so that people using AI tools can end up becoming more biased in domains ranging from social judgments to basic perception.

"Importantly, however, we also found that interacting with accurate AIs can improve people's judgments, so it's vital that AI systems are refined to be as unbiased and as accurate as possible."

Professor Sharot added, "Algorithm developers have a great responsibility in designing AI systems; the influence of AI biases could have profound implications as AI becomes increasingly prevalent in many aspects of our lives."

More information: How human–AI feedback loops alter human perceptual, emotional and social judgements, *Nature Human Behaviour* (2024). DOI: 10.1038/s41562-024-02077-2

Provided by University College London