theconversation.com

Algorithms help people see and correct their biases, study shows

Carey K. Morewedge

6–7 minutes

Algorithms are a staple of modern life. People rely on algorithmic recommendations to wade through deep catalogs and find the best movies, routes, information, products, people and investments. Because people train algorithms on their decisions – for example, algorithms that make recommendations on e-commerce and social media sites – algorithms learn and codify human biases.

Algorithmic recommendations exhibit bias toward popular choices and information that evokes outrage, such as partisan news. At a societal level, algorithmic biases perpetuate and amplify structural racial bias in the judicial system, gender bias in the people companies hire, and wealth inequality in urban development.

Algorithmic bias can also be used to reduce human bias. Algorithms can reveal hidden structural biases in organizations. In a paper published in the Proceedings of the National Academy of Science, my colleagues and I found that algorithmic bias can help people better recognize and correct biases in themselves.

The bias in the mirror

In nine experiments, Begum Celikitutan, Romain Cadario and I had research participants rate Uber drivers or Airbnb listings on their driving skill, trustworthiness or the likelihood that they would rent the listing. We gave participants relevant details, like the number of trips they'd driven, a description of the property, or a star rating. We

also included an irrelevant biasing piece of information: a photograph revealed the age, gender and attractiveness of drivers, or a name that implied that listing hosts were white or Black.

After participants made their ratings, we showed them one of two ratings summaries: one showing their own ratings, or one showing the ratings of an algorithm that was trained on their ratings. We told participants about the biasing feature that might have influenced these ratings; for example, that Airbnb guests are less likely to rent from hosts with distinctly African American names. We then asked them to judge how much influence the bias had on the ratings in the summaries.

The author describes how algorithms can be useful as a mirror of people's biases.

Whether participants assessed the biasing influence of race, age, gender or attractiveness, they saw more bias in ratings made by algorithms than themselves. This algorithmic mirror effect held whether participants judged the ratings of real algorithms or we showed participants their own ratings and deceptively told them that an algorithm made those ratings.

Participants saw more bias in the decisions of algorithms than in their own decisions, even when we gave participants a cash bonus if their bias judgments matched the judgments made by a different participant who saw the same decisions. The algorithmic mirror effect held even if participants were in the marginalized category – for example, by identifying as a woman or as Black.

Research participants were as able to see biases in algorithms trained on their own decisions as they were able to see biases in the decisions of other people. Also, participants were more likely to see the influence of racial bias in the decisions of algorithms than in their own decisions, but they were equally likely to see the influence of defensible features, like star ratings, on the decisions of algorithms and on their own decisions.

Bias blind spot

People see more of their biases in algorithms because the algorithms remove people's bias blind spots. It is easier to see biases in others' decisions than in your own because you use different evidence to evaluate them.

When examining your decisions for bias, you search for evidence of conscious bias – whether you thought about race, gender, age, status or other unwarranted features when deciding. You overlook and excuse bias in your decisions because you lack access to the associative machinery that drives your intuitive judgments, where bias often plays out. You might think, "I didn't think of their race or gender when I hired them. I hired them on merit alone."

The bias blind spot explained.

When examining others' decisions for bias, you lack access to the processes they used to make the decisions. So you examine their decisions for bias, where bias is evident and harder to excuse. You might see, for example, that they only hired white men.

Algorithms remove the bias blind spot because you see algorithms more like you see other people than yourself. The decision-making processes of algorithms are a black box, similar to how other people's thoughts are inaccessible to you.

Participants in our study who were most likely to demonstrate the bias blind spot were most likely to see more bias in the decisions of algorithms than in their own decisions.

People also externalize bias in algorithms. Seeing bias in algorithms is less threatening than seeing bias in yourself, even when algorithms are trained on your choices. People put the blame on algorithms. Algorithms are trained on human decisions, yet people call the reflected bias "algorithmic bias."

Corrective lens

Our experiments show that people are also more likely to correct their biases when they are reflected in algorithms. In a final experiment, we gave participants a chance to correct the ratings

they evaluated. We showed each participant their own ratings, which we attributed either to the participant or to an algorithm trained on their decisions.

Participants were more likely to correct the ratings when they were attributed to an algorithm because they believed the ratings were more biased. As a result, the final corrected ratings were less biased when they were attributed to an algorithm.

Algorithmic biases that have pernicious effects have been well documented. Our findings show that algorithmic bias can be leveraged for good. The first step to correct bias is to recognize its influence and direction. As mirrors revealing our biases, algorithms may improve our decision-making.